THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

AFFECT RECOGNITION BASED ON KEY FRAME SELECTION FROM VIDEO

Master of Science Thesis

MEHMET KAYAOĞLU

ISTANBUL, 2016

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
ELECTRICAL AND ELECTRONICS ENGINEERING

AFFECT RECOGNITION BASED ON KEY FRAME SELECTION FROM VIDEO

Master of Science Thesis

MEHMET KAYAOĞLU

Supervisor: PROF. ÇİĞDEM EROĞLU ERDEM

İSTANBUL, 2016

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
ELECTRICAL AND ELECTRONICS ENGINEERING

Title of the Master's Thesis: Audio Visual Affect Recognition
Name/Last Name of the Student: Mehmet Kayaoğlu
Date of Thesis Defense: 04-01-2016

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assoc. Prof. Nafiz Arıca
Graduate School Director

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Science.

Asst. Prof. Ayça Yalçın Özkumur
Program Coordinator

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Examining Committee Members	Signature
Thesis Supervisor Prof. Çiğdem EROĞLU ERDEM	.....
Member Assoc. Prof. Dr. Hazım Kemal EKENEL	.....
Member Asst. Prof. Dr. Tarkan AYDIN	.....

ABSTRACT

AFFECT RECOGNITION BASED ON KEY FRAME SELECTION FROM VIDEO

Mehmet Kayaoğlu

Department of Electrical and Electronic Engineering

Thesis Supervisor: Prof. Çiğdem Eroğlu Erdem

January 2016, 74 pages

In daily human-to-human interactions, our facial expressions convey non-verbal messages about our emotions and mental states that complement our verbal messages. In the future, human-computer interaction scenarios are also expected to have the ability to recognize emotions to provide more natural man-machine interaction and ubiquitous computing applications such as health care, education, psychology and security.

In this dissertation, we present a multimodal affect recognition method using facial expressions and the speech signal. Given a video with an emotional expression, the frames in the video generally reflect the emotion with different intensities. Moreover, some parts of the video might have little motion, which makes consecutive frames to be very similar to each other. Therefore, we aim to summarize the content of the video by selecting key frames effectively by adopting a recent video summarization method based in minimum sparse reconstruction. We extract static appearance-based features from the selected facial key frames and average them to summarize the visual content of the whole video. We also capture the temporal variations of facial expressions using spatio-temporal appearance based features. Along with visual features, we employ spectral and linear prediction based audio features and fuse them with the video-based features at the score (decision) level. We tested the proposed framework on several databases and also obtained promising results in the ACM International Conference on Multimodal Interaction (ICMI) Emotion Recognition in the Wild (Emotiw 2015) challenge using the proposed method.

Keywords: Affect Recognition, Peak Frame Selection, Affective Computing

ÖZET

VİDEODAN ANAHTAR ÇERÇEVE SEÇİMİNE DAYALI DUYGU TANIMA

Mehmet Kayaoğlu

Elektrik – Elektronik Mühendisliği

Tez Danışmanı: Prof. Çiğdem Eroğlu Erdem

Ocak 2016, 74 sayfa

Günlük yaşantımızda yüz ifadelerimiz duygusal ve zihinsel durumumuz hakkında sözlü olmayan mesajlar taşırlar. Yüz ve ses ifadelerinden duygu tanıma sağlık, eğitim, psikoloji ve güvenlik gibi çok farklı alanlarda kullanılabilmektedir. Yakın gelecekte insan-makine etkileşiminde duygusal durumun daha başarılı olarak tespiti ve buna göre etkileşimin yönlendirilmesi ile daha doğal uygulamaların gerçekleşmesi mümkün olacaktır.

Bu tezde, video dizilerindeki yüz ifadelerini ve konuşma sinyalini kullanarak anahtar video karesi seçimine dayalı duygu tanımaya dayanan bir yöntem öneriyoruz. Duygusal bir ifadenin bulunduğu bir video göz önüne alındığında videoda bulunan her çerçeve genellikle farklı şiddetlerde duygu yansıtmaktadır. Ayrıca videonun bazı bölümlerindeki ardışık karelerin birbirine çok benzer olmasından dolayı yüzde küçük hareketler olmaktadır. Etkili anahtar çerçeve seçimiyle tüm videoyu en az çerçeve ile ve en etkili biçimde özetlemeyi hedefledik. Bunun için en az seyrek geriçatıma dayalı bir yöntem kullandık. Seçilen anahtar çerçevelere ait özniteliklerin ortalamasını alarak tüm videoya ait duygu içeriğini temsil etmek için kullandık. Ayrıca videodaki zamansal değişimleri de değerlendirmek için zamansal-uzamsal özniteliklerden yararlandık. Görsel özelliklerin yanında ses verisine ait spektral ve doğrusal kestirime dayalı öznitelikleri kullanarak görsel duygu tanımaya sonuç seviyesinde birleştirdik. Önerdiğimiz sistemi çeşitli veri tabanları üzerinde denedik ve önerilen bu sistemle ACM International Conference on Multimodal Interaction (ICMI) Emotion Recognition in the Wild (Emotiw 2015) yarışmasına katılarak olumlu sonuçlar elde ettik.

Anahtar Kelimeler: Duygu Tanıma, Anahtar Kelime Seçme, Duygusal Hesaplama

CONTENTS

TABLES vii

FIGURES ix

ABBREVIATIONS xi

1. INTRODUCTION 1

1.1 MOTIVATION 1

1.2 PROBLEM STATEMENT 2

1.3 CONTRIBUTIONS AND ORGANISATION OF THE THESIS 3

2. LITERATURE SURVEY 5

2.1 HUMAN AFFECT (EMOTION) PERCEPTION 5

2.1.1 The Description of Emotion and Affect 5

2.1.2 Association between Affect, Audio, and Visual Signals 6

2.2 EXISTING EMOTIONAL DATABASES 6

2.3 AUDIO-BASED AFFECT RECOGNITION 9

2.3.1 Emotion Related Speech Features 10

2.3.2 Audio-Based Affect Recognition Studies 11

2.4 VISION-BASED AFFECT RECOGNITION 15

2.4.1 Face Registration 17

2.4.1.1 Whole face registration 17

2.4.1.2 Part and point based registrations 17

2.4.2 Feature Representations 18

2.4.3 Facial Emotion Recognition Studies 20

2.5 AUDIO-VISUAL AFFECT RECOGNITION 22

2.5.1 Multimodal Fusion Methods 22

2.5.1.1 Feature level fusion 22

2.5.1.2 Score level fusion 22

2.5.1.2.1 Minimum rule 23

2.5.1.2.2 Maximum Rule 23

2.5.1.2.3 Sum Rule 23

2.5.1.2.4 Mean Rule 23

2.5.1.2.5 Weighted Average Rule 23

2.5.1.2.6 Product Rule 24

2.5.1.2.7 Bayesian Framework 24

2.5.2 Multimodal Emotion Recognition Studies 24
3. AFFECT RECOGNITION USING KEY FRAME SELECTION BASED ON MINIMUM SPARSE RECONSTRUCTION 27
3.1 PREPROCESSING OF FACE IMAGES 27
3.2 EXTRACTION OF VISUAL FEATURES FROM FACE IMAGES 28
3.2.1 Local Phase Quantization Features 28
3.2.2 LBP TOP Features 31
3.3 EXTRACTION OF AUDIO FEATURES 33
3.3.1 Mel Frequency Cepstral Coefficients (MFCC) Features 33
3.3.2 Relative Spectral Transform – Perceptual Linear Prediction (RASTA - PLP) Features 34
3.4 KEY FRAME SELECTION BASED ON VIDEO SUMMARIZATION 35
3.5 CLASSIFICATION 39
4. EXPERIMENTAL RESULTS 42
4.1 DATABASES USED 42
4.2 EXPERIMENTAL SETUPS AND RESULTS 42
4.2.1 Results on CK+ Database 42
4.2.2 Results on eNTERFACE'05 Database 48
4.2.3 Results on BAUM-1a Database 52
4.2.4 Emotion Recognition in the Wild (EmotiW 2015) Challenge 57
5. CONCLUSION AND FUTURE WORK 62
5.1 CONCLUSION 62
5.2 FURTHER RESEARCH 63
REFERENCES 64

TABLES

Table 2.1: Audio / Visual databases of human affective behavior 9
Table 2.2: A list of representative works of audio based emotion recognition..... 14
Table 2.3: Popular methods that are being used on facial expression recognition. 19
Table 2.4: A list of representative works in the field of video emotion recognition 21
Table 4.1: Number of frames for CK+ sequences..... 43
Table 4.2: Statistics of selected keyframe counts for CK+ sequences..... 45
Table 4.3: Facial recognition rates for the CK+ database..... 46
Table 4.4: Confusion matrix for the CK+ database when the last frame (i.e. the peak frame) and CK+ landmarks are used 47
Table 4.5: Confusion matrix for the CK+ database using the selected key frames and CK+ landmarks..... 47
Table 4.6: Performances of current facial-expression recognition methods on CK+ 48
Table 4.7: Frame counts for eNTERFACE'05 sequences and selected key frames 48
Table 4.8: Recognition rates for eNTERFACE'05 database 50
Table 4.9: Confusion matrix for eNTERFACE'05 database when LPQ of selected key frames are used 51
Table 4.10: Performances of expression recognition methods on eNTERFACE'05..... 51
Table 4.11: Peak frame selection methods on eNTERFACE'05 database. 52
Table 4.12: Properties of the BAUM-1 database 53
Table 4.13: Actual and extracted number of key frames for BAUM-1a 5 emotion case 54
Table 4.14: Recognition rates for the BAUM-1a database for 5 emotions..... 55
Table 4.15: Confusion matrix for 5 basic emotions using BAUM-1a database when LPQ of selected key frames are used. 55
Table 4.16: Frame counts for BAUM-1a 8 emotion case 55
Table 4.17: Recognition rates for the BAUM-1a 8 emotion..... 56
Table 4.18: Confusion matrix for the 8 basic emotions using BAUM-1a when LPQ of selected key frames are used..... 56
Table 4.19: Peak frame selection methods on BAUM-1a..... 57
Table 4.20: The numbers of samples in EmotiW 2015 AFEW database..... 58
Table 4.21: Accuracies of the 8 test cases on validation and test sets. 59

Table 4.22: Confusion matrix of Case 8 (audio-visual) on validation set. 60
Table 4.23: Confusion matrix of Case 8 (audio-visual) on test set. 60
Table 4.24: Confusion matrix of Case 2 (Visual) on test set. 60
Table 4.25: Confusion matrix of Case 6 (Audio) on test set. 61

FIGURES

Figure 1.1: Modalities of emotion recognition 1
Figure 1.2: Potential applications of affect recognition 2
Figure 2.1: Prototypical six basic emotions according to Ekman. 5
Figure 2.2: Emotion recognition from speech signal 10
Figure 2.3: Block Diagram of PLP Processing 11
Figure 2.4: Muscles of head and neck 15
Figure 2.5: Overview of Facial Action Coding System 16
Figure 2.6: Basic structure of a facial expression recognition system. 16
Figure 3.1: General emotion recognition framework 27
Figure 3.2: Cropping and resizing of an image, based on eye center locations 28
Figure 3.3: Sub region selection for feature extraction 31
Figure 3.4: LPQ feature extraction from selected sub regions 31
Figure 3.5: The texture of LBP-TOP planes and the corresponding histograms 32
Figure 3.6: The structure of MFCCs 34
Figure 3.7: The structure of RASTA-PLP Method 34
Figure 3.8: Several sequences from CK+ database containing single emotion 35
Figure 3.9: General framework of peak frame selected affect recognition system 36
Figure 3.10: An example of video summarization from AFEW 5.0 database. 38
Figure 3.11: An example of video summarization from CK+ database. 39
Figure 3.12: Binary linear classifier. 40
Figure 3.13: Fusion of LBP-TOP and LPQ video features 41
Figure 3.14: Fusion of LBP-TOP, LPQ and OpenSmile audio features 41
Figure 4.1: Expression intensity among frame sequences of CK+ Database 42
Figure 4.2: The 68 landmarks given in the CK+ database 43
Figure 4.3: Number of frames in CK+ sequences 43
Figure 4.4: Key frame extraction when two frames are selected. 44
Figure 4.5: Key frame extraction when only one (last) frame is selected. 44
Figure 4.6: Selected key frames are marked with red star for all CK+ sequences 45
Figure 4.7: Number of selected keyframes for CK+ sequences 45
Figure 4.8: A sample sequence from eNTERFACE'05 database 48

Figure 4.9: Distribution of the number of frames for eNTERFACE'05 sequences.....49
Figure 4.10: Number of frames in eNTERFACE'05 sequences.....49
Figure 4.11: Key frame locations are marked for BAUM-1a database for 5 emotions..54
Figure 4.12: Frame counts of BAUM-1a 8 emotion.....56
Figure 4.13: Noisy face detection and tracking results from AFEW 5.0 dataset.....58
Figure 4.14: EmotiW 2015 challenge results on test sets 61

ABBREVIATIONS

AAM : Active Appearance Modem
AFEW : Acted Facial Expressions in the Wild
ANN : Artificial Neural Network
ASM : Active Shape Model
AU : Action Unit
BAUM : Bahçeşehir University Multimodal Affective Database
BOW : Bag-of-Words Representation
CK+ : Extended Cohn-Kanade Database
DBN : Dynamic Bayesian Network
DCT : Discrete Cosine Transform
DFT : Discrete Fourier Transform
EMO-DB : Berlin Emotional Speech Database
EmotiW : Emotion Recognition in the Wild
ENIAC : Electronic Numerical Integrator and Computer
FACS : Facial Action Coding System
FAP : Facial Animation Parameter
FFT : Fast Fourier Transformation
GMM : Gaussian Mixture Model
HCI : Human Computer Interaction
HCRF : Hidden Conditional Random Field
HOG : Histogram Of Oriented Gradients
HMM : Hidden Markov Models
JAFFE : Japanese Female Facial Expression
KDEF : Karolinska Directed Emotional Faces
K-NN : K-Nearest Neighbors
LBP : Local Binary Patters
LBP-TOP : Local Binary Pattern Histograms from Three Orthogonal Planes
LDC : Linear Discriminant Classification
LOSO : Leave-One Subject-Out
LPQ : Local Phase Quantization
LSRE : Least Square Reconstruction Error

MFCC	: Mel-frequency Cepstral Coefficients
MLP	: Multilayer Perceptron
NN	: Neural Network
OSP	: Orthogonal Subspace Projection
POR	: Percentage of Reconstruction Error
PIE	: Pose, Illumination, And Expression
PLP	: Perceptual Linear Predictive
PSF	: Point Spread Function
RASTA-PLP	: Relative Spectral Transform - Perceptual Linear Prediction
SIFT	: Scale Invariant Feature Transform
SVM	: Support Vector Machines
TAN	: Tree-Augmented Naive Bayes

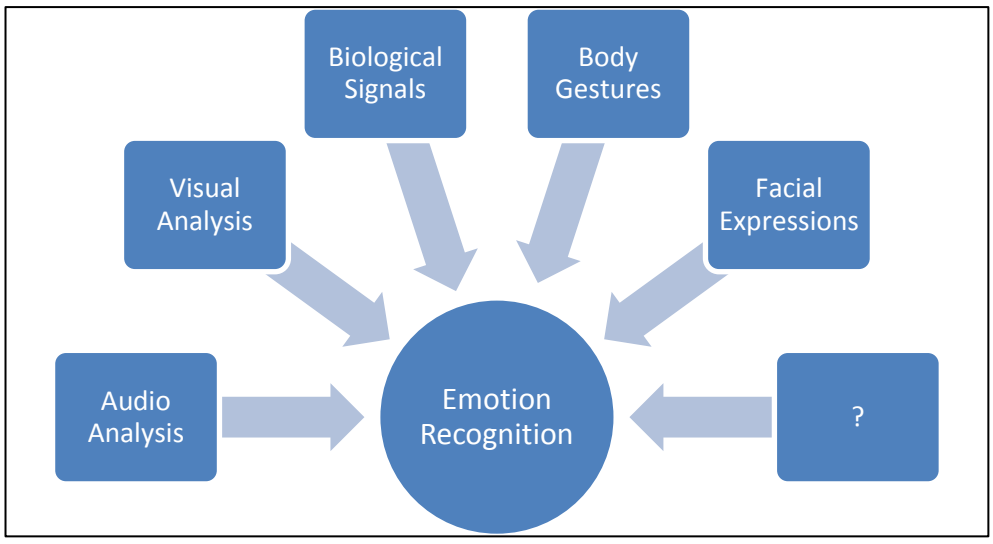
1. INTRODUCTION

1.1 MOTIVATION

Early computers such as ENIAC (Electronic Numerical Integrator And Computer) were only basic tools that solved basic numerical problems and it was not so long ago. However, by the availability of high computing power, computers are being utilized in many aspects of our daily lives. Therefore, the interaction between humans and computers are gaining increasing importance every day in many application areas. In the light of these developments, human computer interaction (HCI) has become one of the most important research areas.

During a successful human-to-human or human-to-computer interaction, recognition of emotions has a crucial role. Although humans mainly express their emotions via linguistic communication which is based on words and sentences, there are several other channels that are also being used to express emotions. Face and body gestures can be considered as second important and mostly used visual instruments for affect representation and recognition. For example, smile may be the unavoidable outcome of happiness and similarly frown may be an inevitable sign of anger.

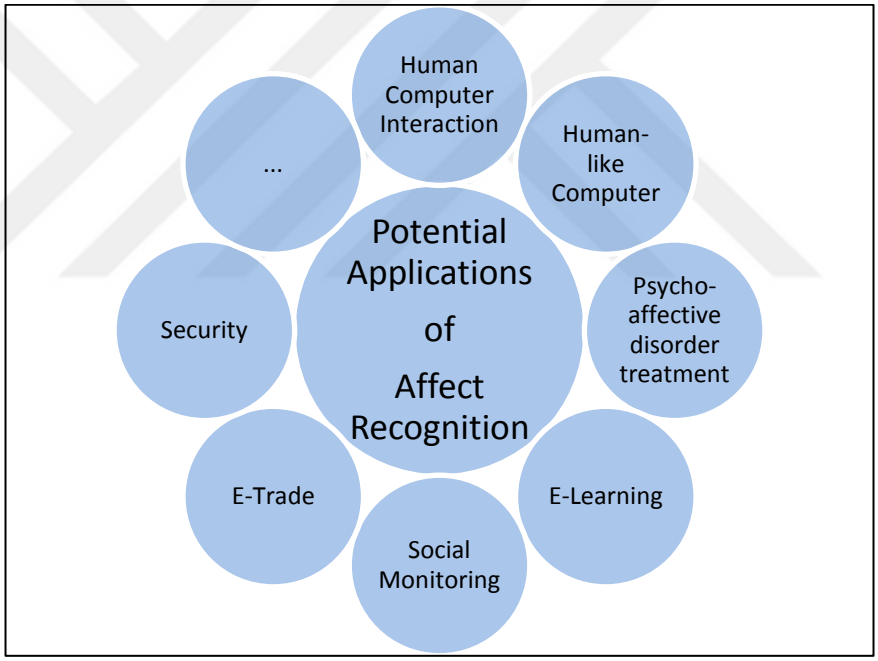
Figure 1.1: Modalities of emotion recognition



Designing human-like computers is an interesting yet challenging problem for many application areas, including making advertisements by considering the emotional state

of a person [134], treatment of psycho-affective illnesses [117], health care [77] and security [108]. For instance, in e-learning applications, affect recognition can be used for increasing the efficiency of the overall teaching system by detecting student's emotional and/or mental state (interested or bored). If the student is interested, new material may be added, or if he is confused, the material may be explained from different perspective or using different examples. Also emotion recognition may be used in digital pets and robots to react according to user's affective state. Inevitably, this will increase the realism of robot's action. Affect recognition may also help the treatment of psychological health problems, when the patient is not capable of expressing himself verbally.

Figure 1.2: Potential applications of affect recognition



1.2 PROBLEM STATEMENT

Visual signals are one of the main modalities which are being used for expressing emotions. Affect recognition systems may use images or videos of facial expressions to capture the visual signals expressed by the face. An image reflects a 'snapshot' of an emotional facial expression, whereas a video also reflects the temporal evolution of the emotion. Therefore, when a video includes an emotional facial expression, the frames in the video generally reflect the emotion with different intensities. Moreover, some parts

of the video might have little motion, which makes subsequent frames to be very similar to each other. Hence it may not be necessary to process all the frames in a given affective face video. Therefore, we aim to recognize the affect in a video using a method based on video summarization by key frame selection. The key frame selection uses a minimum sparse reconstruction approach with the goal of representing the original video using the least number of "key" frames in the "best possible" way. We also utilize the audio channel for emotion recognition and fuse it with the visual channel at the decision level.

1.3 CONTRIBUTIONS AND ORGANISATION OF THE THESIS

In this thesis, we present multi-modal emotion recognition system based on fusion of facial expressions and the speech modalities. In order to recognize the emotions from facial expressions, we adopt a key frame selection method using video summarization based on minimum sparse reconstruction. The presented key frame selection method is tested and evaluated on three different databases, namely, CK+, eNTERFACE'05 and BAUM1a databases, which showed that the method outperforms previous key frame selection methods in the literature [140][141]. We also participated in the ACM International Conference on Multimodal Interaction (ICMI) Emotion Recognition in the Wild (Emotiw 2015) challenge, which based on the AFEW database. We achieved an audio-visual classification accuracy of 49.91 percent on the test set while the video based base-line accuracy was 39.13 percent. Our audio-visual affect recognition accuracy was the 7th among 13 teams, which participated in the challenge. We also published the below paper at ACM International Conference on Multimodal Interaction (EmotiW 2015 challenge);

- M. Kayaoglu, C. E. Erdem, "Affect Recognition using Key Frame Selection based on Minimum Sparse Reconstruction", *ACM Int. Conf. on Multimodal Interaction (*ICMI*), The Third Emotion Recognition in the Wild Challenge (EmotiW 2015)*, November 9-13, Seattle, USA, 2015. DOI: http://dx.doi.org/10.1145/2818346.2830594

The organization of the thesis is as follows. In Chapter 2, a brief literature survey on emotion recognition methods is presented. Chapter 3 introduces our audio-visual affect recognition framework and the minimum sparse reconstruction based key frame

selection method. In Chapter 4, experimental results on four different databases are presented. Finally, conclusions and future directions for research are presented in Chapter 5.

2. LITERATURE SURVEY

2.1 HUMAN AFFECT (EMOTION) PERCEPTION

First published work on emotion recognition can be considered as Darwin's publication [30] in 1872. In his work Darwin states that emotions were initiated in certain situations and common for all humans and animals. In the light of Darwin's view, Ekman defined and presented [42] six basic emotions as anger, disgust, fear, happiness, sadness and surprise in 1999 and showed that they are universal(i.e. culture independent). Figure 2.1 shows examples of Ekman's six basic emotions.

Figure 2.1: Prototypical six basic emotions according to Ekman.
From left to right: Anger, Fear, Disgust, Surprise, Happiness, and Sadness.

Source: https://medium.com/@iheartliterati/emotions-expressions-and-signals-916eb406c2f8

Scherer tests in 1998 [50] showed that even in congenitally blind individuals, facial expressions are similar to those without this disability. This exploration has motivated the studies on automatic emotion recognition that started by Ekman's defining basic emotions in 1992.

2.1.1 The Description of Emotion and Affect

In our everyday lives, emotions reflect a person's state of mind and instinctive responses. It is a state of feeling that results in physical and psychological changes that influence our behavior. Affect is a psychological term for an observable expression of emotion. Common examples of affect are sadness, fear, happiness, anger.

Emotion and affect are hard to comprehend and currently, there is no consensus about the exact definition of them. A computer scientist would definitely have a different definition from that of a psychologist, behavioral scientist or an average person.

2.1.2 Association between Affect, Audio, and Visual Signals

As it is mentioned in [90], a person's affect is the expression of emotion or feelings displayed to others through facial expressions, hand and body gestures, voice tone, and other emotional signs such as laughter, tears, heartbeat or body temperature. Each of these biological signs used for expressing emotions is called a mode.

What is considered a normal range of affect, called the broad affect, varies from culture to culture, and even within a culture. Certain individuals may gesture prolifically while talking, and display dramatic facial expressions in reaction to social situations or other stimuli. Others may show little outward response to social environments or interactions, expressing a narrow range of emotions to the outside world.

2.2 EXISTING EMOTIONAL DATABASES

Accessing fully labeled databases prior to development and implementation of an emotion recognition system plays crucial role for researchers working in this field. Using publicly available databases is one of the solutions. However, recognizing the emotion of a person is not an easy task even for human observers. In order to get emotionally labeled data, different approaches are used in practice. Although the most ideal approach is capturing natural emotional expressions, it is not easy to collect this kind of data without influencing the subject. So, available databases in the literature contain mostly acted (simulated), and sometimes elicited emotional expressions.

The acted emotions are obtained by asking the subjects to act a predefined emotion. The subjects are usually professional actors/actresses since they can portray emotions more realistically. However, the acted emotions are often exaggerated compared to those expressed naturally. A way of partially overcoming this drawback is to use emotion-triggering texts and/or simulations to evoke the subject's emotion. This type of databases are called elicited emotional databases and are better than the acted ones in the sense of being more realistic. Ideally, natural databases are a better choice for emotion recognition systems since they contain naturalistic and unbiased emotions [37]. Below we briefly review the most widely used affective databases in the literature.

Japanese Female Facial Expression (JAFFE) database [83] is one of the first emotional databases. It contains 213 images of 7 facial expressions (6 basic facial expressions and 1 neutral) posed by 10 Japanese female models. Each image has been

rated on 6 emotion adjectives by 60 Japanese subjects. The photos were taken at the Psychology Department in Kyushu University.

Karolinska Directed Emotional Faces (KDEF) database [82] contains 4900 images of 70 individuals with an age range between 20 and 30 years. Each subject is displaying 7 different emotional expressions and each expression being photographed (twice) from 5 different angles.

The Pose, Illumination, and Expression (PIE) database [116] contains 41368 images of 68 people from 13 different poses, in 43 different illumination conditions, and with 4 different expressions. The database is collected between October and December 2000.

The **Cohn-Kanade Facial Expression database** [66] contains sequences of images starting with a neutral expression and ending at the target emotion. There are a total of 327 sequences with emotion labels from 123 subjects in the extended version of the database (CK+) [79].

The FEEDTUM database [132] contains spontaneous video clips with elicited emotions recorded from 18 subjects with six basic emotions and the neutral expression.

The MMI database [102][124] is conceived in 2002 and it consists of over 2900 videos and high-resolution still images of 75 subjects displaying the six basic emotions. It consists of both acted and spontaneous expressions. It is fully annotated for the presence of AUs in videos and partially coded on frame-level, indicating for each frame whether an AU is in either the neutral, onset, apex or offset phase.

The Berlin Emotional Speech database (EMO-DB) [15] is one of the emotional speech databases available to researchers and it contains 495 samples of acted emotional speech in German language. This database is comprised of 10 different texts and 10 different actors in a happy, angry, anxious, fearful, bored and disgusted way as well as in a neutral version.

The AIBO database [9][10] contains 9 hours spontaneous, in German, emotional reactions of 51 children at the age 10-13 years interacting with Sony's pet robot Aibo. The data is annotated with 11 emotion categories by five human labelers on the word level.

The BU-3DFE database of Yin et al. [74] contains 3D range data of six basic facial expressions displayed at four different intensity level.

The FABO database of Gunes and Piccardi [54] contains videos of facial expressions and body gestures that belong to uncertainty, anxiety, boredom, and neutral emotions besides six basic emotions.

One of the early audio-visual databases is the **Belfast database** [118] which contains 298 audio-visual clips extracted from television talk shows, current affairs programs and interviews conducted by the research team from 125 English speaking subjects, 31 male, 94 female. Clips from the first 100 speakers, totaling 86 min of speech, have been labeled psychologically and acoustically. It contains a wide range of emotions.

The HUMAINE database [39] is a combination of 50 clips from Belfast and some other databases containing both spontaneous and acted data.

The Bahcesehir Universiy Multimodal Affective Database - 1 (BAUM-1) [98] is another database collected from 31 subjects containing all 6 basic emotions along with contempt, boredom and some mental states. It is a collection of audio-visual facial clips of 280 acted and 1222 spontaneous (re-acted) affective expressions. The audio-visual clips are in Turkish.

The Bahcesehir Universiy Multimodal Affective Database - 2 (BAUM-2) [46] is a dataset of audio-visual affective facial clips extracted semi-automatically from movies and TV series. It includes 1047 video clips. The BAUM-2 database consists of clips in two languages, 616 of which are in English and 431 of which are in Turkish. It contains facial clips with various head poses, illumination conditions, occlusions, and subjects from various ages, and races recorded under close-to-natural conditions.

The eNTERFACE'05 database [88] contains acted emotional recordings of 34 men and 8 women subjects of 14 different nationalities speaking in English and showing prototypical emotions. Six basic emotional states are expressed in the video clips of the database in an acted way by listening short stories and eliciting a particular emotion.

Early databases used were mostly acted and recorded in laboratory conditions under controlled head pose and illumination variations. Recently, more spontaneous and close to real world databases have been collected [34], [38], [46], some of which have been used in challenges such as FERA 2015 [125], and AVEC 2014 [127], and Emotion Recognition in the Wild (EmotiW 2015) challenge [35].

The EmotiW challenge consists of categorical audio-video based emotion recognition based on the **Acted Facial Expression in Wild database (AFEW 5.0)**. The AFEW

database contains short audio-visual clips collected from movies and labeled using a semi-automatic approach described in [34]. This challenge is a continuation of the EmotiW 2013 and 2014 challenges and the task is to assign a single emotion label to the video clip from the seven emotions (Anger, Disgust, Fear, Happiness, Neutral, Sad and Surprise). The AFEW database is quite challenging since the video clips contain variability in illumination and head pose, as well as severe occlusion and complex background. All aforementioned databases are summarized in Table 2.1.

Table 2.1: Audio / Visual databases of human affective behavior

Name	A/V	# Subjects	Language	BE	nBE	Posed/Not-Posed
EMO-DB	A	10	German	6+N		P
AIBO	A	51	German	-	Var	Elicited
JAFFE	V	10	-	6+N		P
KDEF	V	70	-	6+N		P
PIE	V	68	-	-	4	P
FEEDTUM	V	18	-	6+N		nP
CK	V	97	-	6		P
CK+	V	327	-	6		P + nP
MMI	AV	75	-	6+N		P + nP
UT Dallas '06	V	229			6 BE + Var	nP
Belfast Naturalistic	AV	125	English	4+N	4 BE + N	Naturalistic
Belfast Induced	AV	256	English	6+N		
HUMAINE	AV	18	English French Hebrew	-	-	P + nP
Enterface'05	AV	42	English	-	6 BE	P
BAUM-1	AV	31	Turkish	-	Var	nP
BAUM-2	AV	286	Turkish English	6+N	-	nP
BU-3DFE	V	100	-	6		P
FABO	V	23	-	-	6 BE + Var	P
GEMEP	-	10	-	6+N	12	P
SEMAINE	-	150	-	3	10	nP
DISFA	-	27	-	-	-	nP

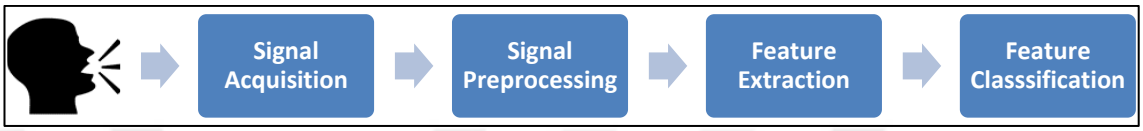
A: Audio, V: Visual, BE: Basic Emotions, nBE: Non-Basic Emotions, N: Neutral, Var: Various non-basic emotions

2.3 AUDIO-BASED AFFECT RECOGNITION

Speech is one of the indispensable means for sharing ideas, observations, and feelings. People usually convey emotions by using speech information either explicitly through linguistic or implicitly through paralinguistic messages. Hence, considering only the verbal part, without taking into account the manner in which it was spoken, will cause loss of important aspects of the spoken message.

There are many application areas of speech emotion recognition. For example, speech recognition systems need to analyze the speech correctly in order to perform effectively under changes in emotions, states and tone of speakers. Doctors can diagnose possible diseases in patients. Psychologist can predict the human state of mind and human-computer interaction experts can enrich the communication between human and machines.

Figure 2.2: Emotion recognition from speech signal



The general emotion recognition process from speech signals can be divided into the following steps: (1) capturing the speech signal and preprocessing, (2) extracting audio features, (3) recognition of emotions with an appropriate classifier. Preprocessing step of speech can include detection of voiced and unvoiced segments, end point detection, dividing signals into frames with predefined length and windowing them. Figure 2.2 shows general architecture of the emotion recognition system from speech signal.

2.3.1 Emotion Related Speech Features

Emotion related speech features can be considered in two categories, prosodic and spectral features. Prosody is the study of the tune and rhythm of speech and how these features contribute to meaning. Prosody can be characterized by vocal pitch (fundamental frequency) [64], loudness (energy) [85] and temporal components (rhythm, duration) [138]. So **prosodic features** are based on time domain related features of the speech and generally simple to extract and have easy physical interpretation, like: the energy of signal, zero crossing rate, maximum amplitude, minimum energy, etc. Human beings use long term speech related prosodic features to perceive the emotional content from speech. Therefore, prosodic features are commonly used for automatic recognition of emotions from speech.

The **spectral features** are looking for the frequency related characteristics of the speech signal. These features are obtained by converting the time based signal into the frequency domain using the Fourier Transform. The basic spectral features are the

fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off and etc. [146]. These features can be used to identify the notes, pitch, rhythm, and melody.

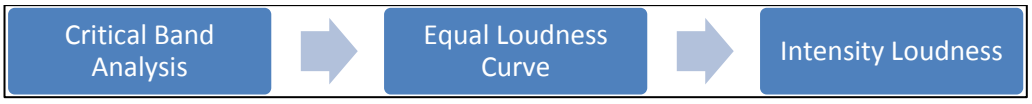
The most popular spectral speech feature representation currently used is Mel-frequency Cepstral Coefficients (MFCC). They were introduced by Davis and Mermelstein in the beginning of 1980's [32], and have been widely used ever since [111][92][21][17].

High level implementation steps of MFCC can be listed as;

i. Apply windowing to the signal
ii. For each windowed signal, discrete fourier transform (DFT) is applied to get the power spectrum.
iii. Apply the MEL filter bank to the power spectrum and sum the energy in each filter.
iv. Take the logarithm of all filter bank energies.
v. Take the discrete cosine transform (DCT) of the log filter bank energies.
vi. Keep DCT coefficients

Another spectral speech feature which is widely used in emotion recognition is perceptual linear prediction (PLP) [56]. It is known that PLP is good under noisy conditions. As a result of discarding irrelevant information of the speech and transforming spectral characteristics to match human auditory system characteristics, PLP improves the speech recognition rate. Figure 2.3 shows the main steps of PLP computation.

Figure 2.3: Block Diagram of PLP Processing



2.3.2 Audio-Based Affect Recognition Studies

Ang et al. [5], in 2002, explored speech-based recognition of annoyance and frustration. In addition to the prosodic features, they investigated language models and predicted whether an utterance is neutral or frustrated.

In 2003, Schuller et al. [110] introduced emotion recognition by use of continuous hidden Markov models. They introduced two approaches. As a first approach, a

statistical framework of an expression is classified by Gaussian Mixture Models (GMM) using the energy contour and raw pitch of the speech signal. Second proposed method introduced an increased temporal complexity applying continuous hidden Markov models by considering several states using low-level instantaneous features instead of global statistics.

Batliner et al. [11], proposed a module for detecting trouble in communication in which a prosodic classifier is combined with other knowledge sources, such as conversationally peculiar linguistic behavior.

Zhang et al. [143], in 2004, investigated speech-based analysis of confidence, puzzle, and hesitation by means of lexical, prosodic, spectral, and syntactic analyses of users' speech.

In 2005, Steidl et al. [119], proposed a new entropy-based method on the detection of empathy. Hirschberg et al. [57] and Graciarena et al. [51], attempted to distinguish deceptive from non-deceptive speech using machine learning techniques on features extracted from a large corpus of deceptive and non-deceptive speech. Liscombe et al. [75] proposed that acoustic-prosodic features can distinguish certainness from other states. Kwon et al. [69] selected pitch, log energy, formant, mel-band energies, and mel frequency cepstral coefficients (MFCCs) as the base features, and added acceleration of pitch to distinguish stressed versus neutral speech.

Lee et al. [70] proposed that combination of acoustic and language information gives better affect recognition results than focusing on only the acoustic information contained in speech. They used combination of three sources of information (acoustic, lexical, and discourse) for emotion recognition. Their case study was detecting negative and non-negative emotions using spoken language data obtained from a call center application.

Vogt et al. [129] conducted a data-mining experiment on feature selection for automatic emotion recognition. More than 1000 features were derived from pitch, energy and MFCC time series. Then the most relevant features were detected by correlating each feature.

Lee and Narayan [71] explored domain specific emotion recognition from speech signals on a case study of detecting negative and non-negative emotions. They

combined acoustic, lexical, and discourse information and then used Linear Discriminant Classification (LDC) with k-nearest neighbors (K-NN) classifiers.

In 2009, Xia et al. [87] proposed a method based on a hybrid of hidden Markov models (HMMs) and artificial neural network (ANN). In the proposed method, the utterance is viewed as a series of voiced segments, and feature vectors extracted from the segments are normalized into fixed coefficients using orthogonal polynomials methods, and then, distortions are calculated as an input of ANN. Also the utterance as a whole is modeled by HMMs, and likelihood probabilities derived from the HMMs are normalized to be another input of ANN.

In 2010, Erdem et al. [45] proposed a Random Sampling Consensus (RANSAC) based training approach for the problem of emotion recognition from speech. They inserted a data cleaning process to the training phase of the Hidden Markov Models (HMMs) for the purpose of removing suspicious instances of labels that may exist in the dataset.

Table 2.2: A list of representative works of audio based emotion recognition.

Researcher(s)	Features	Classification	Database	Accuracy
Ang et al. [5]	• Pitch • Energy	• Decision trees	• [131]	80.2 %
Schuller et al. [110]	• Pitch • Energy	• GMM	• Self-defined	86 %
Lee et al. [70]	• Pitch • Energy	• LDC	• Real users	77 %
Vogt et al. [129]	• Pitch • Energy • MFCC	• Naive Bayes	• Emo-DB • SmartKom	• 77.4 % • 38.7 %
Lee et al. [71]	• Acoustic features • Discourse information	• LDC	• Real users	F: 40 % M: 36.4 %
Xia et al. [87]	• Pitch • Energy • Formant • LPCC • MFCC	• HMM • ANN	• BHUDES • Emo-DB	• 81.7 % • 71.7 %
Ntalampiras et al. [95]	• Temporal features	• HMM	• Emo-DB	91 %
Erdem et al. [45]	• MFCC • Temporal Features • RANSAC	• HMM	• FAU-Aibo.	41.66 %
Bozkurt et al. [13]	• MFCC • LSF • Temporal Features	• GMM	• EMO-DB • FAU Aibo	• 84.58 % • 40.76 %
Bozkurt et al. [14]	• MFCC • LSF	• HMM	• FAU Aibo	43.59 %
Cheng et al. [24]	• Pitch • MFCC	• GMM	• Self-defined	F: 79.9 % M: 89 %

F: Female, M: Male

In 2010, Bozkurt et al. [13] proposed the use of the line spectral frequency (LSF) features for emotion recognition from speech and then, in 2011[14], used weighted mel-frequency cepstral coefficient (WMFCC) features. They evaluated the WMFCC features together with the standard spectral and prosody features using HMM based classifiers on FAU Aibo emotional speech database. They resulted that unimodal classifiers with the WMFCC features perform better than the classifiers with standard spectral features.

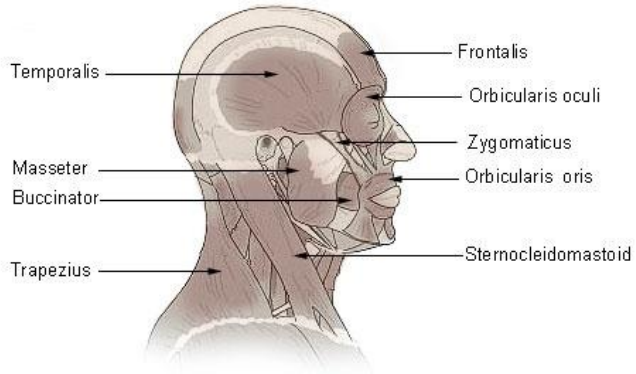
In 2012, Ntalampiras et al. [95] examined short-term statistics, spectral moments, and autoregressive models for speech emotion recognition. They experimented on fusing these sets on the feature and log-likelihood levels based on HMM classification.

Cheng et al. [24] studied on long-term and short-term features of speech. In order to reduce the halving and the doubling errors in pitch tracking, they proposed an algorithm based on the wavelet analysis.

2.4 VISION-BASED AFFECT RECOGNITION

The most commonly used and prominent work for labeling of facial expression is the Facial Action Coding System (FACS) which was conducted by Paul Ekman and Wallace V. Friesen and published in 1978 [41]. Ekman, Friesen, and Joseph C. Hager published a significant update to FACS in 2002 [43].

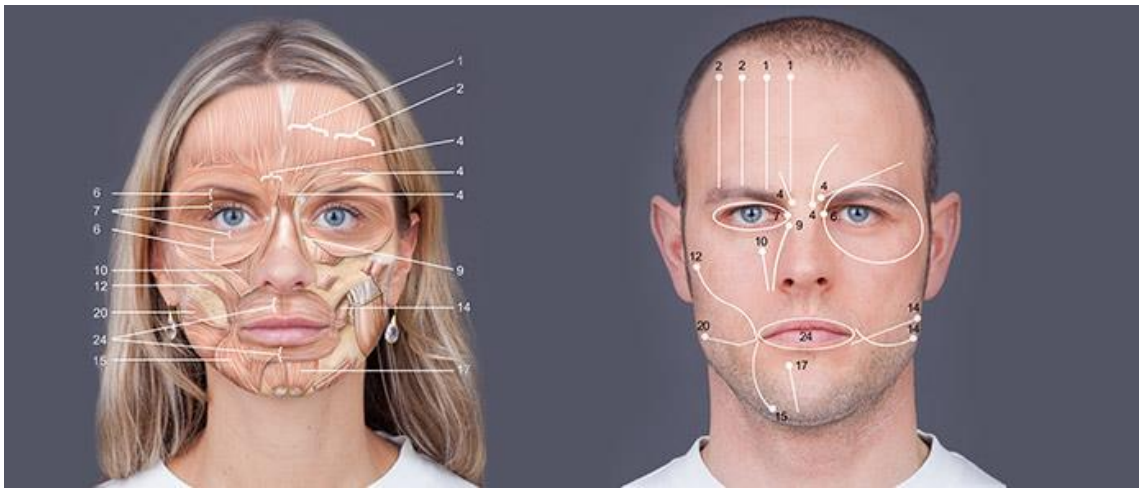
Figure 2.4: Muscles of head and neck



Source: Wikipedia [148].

Ekman stated that facial muscles generate visually perceivable changes in face. So he defined facial changes based on the movements of facial muscles in terms of some facial action units, called as Action Units (AUs). Each facial expression may be described by an individual AU or a group of AUs. The facial muscles and the direction of motion can be seen in Figure 2.5.

Figure 2.5: Overview of Facial Action Coding System

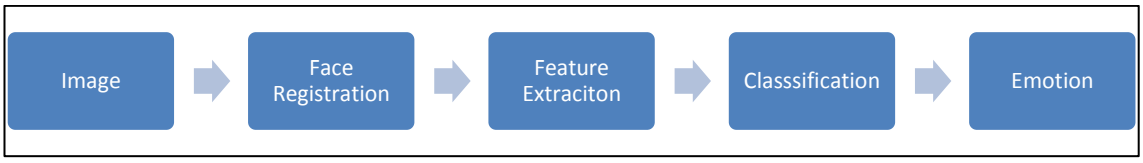


Source: http://www.mimik-lesen.com [149].

Prior to the introduction of FACS, most of the facial behavior research relied on human observation which was not reliable and accurate. Ekman's work on studying the activities of the muscles inspired many researchers for facial expression recognition by means of image and video.

By considering the overall process, every facial expression recognition system is formed by some fundamental components: face registration, feature extraction and classification. These components must be fulfilled in order to classify the expression into a particular emotion. Figure 2.6 summarizes the basic components of a facial expression recognition system.

Figure 2.6: Basic structure of a facial expression recognition system.



In order to detect the facial expression in a given image or video, the first step is face detection. Once the face is detected, it needs to be tracked over time in a video or sequence of frames. Although detection and tracking phases are logically in a sequence, they are tightly bounded together. Most of the methods handle them together in time sequenced frames. So face registration can be thought as a whole combination of face detection and tracking.

After the face is detected, the recognition system should extract some features that represent emotions. Classifying extracted features into a set of emotion classes is the final step to labeling the frame or sequence.

2.4.1 Face Registration

Face registration is initial and crucial step for a facial affect recognition system. Face registration techniques can be divided into three categories: whole face, part base and point based registration.

2.4.1.1 Whole face registration

Some registration techniques are based on locations of the facial points and performed by detecting facial landmarks. Using positions of the landmarks, a global geometric transformation is calculated to transform the input face to a prototypical face.

Although some systems use two eye points or the eyes and nose ([63], [77]), some systems may use more points (e.g. 60-70 points) to compute the transformation [26]. Computing the transformation from more points has some advantages. First, the transformation becomes less sensitive to the registration errors of individual landmark points. Second, the transformation can cope with head-pose variations better, as the facial geometry is captured in more detail.

While rigid registration approaches register the face as a whole object, non-rigid approaches enable registration regionally. Therefore registration errors due to facial activity can be somewhat suppressed. For instance, Active Appearance Models (e.g. [80]) and SIFT-flow [18] can be used for non-rigid registrations.

2.4.1.2 Part and point based registrations

All human faces share some similar properties. For instance, the eyes region is darker than the upper-cheeks and the nose bridge region is brighter than the eyes. Some face registration algorithms process faces by considering these regions like eyes, eyebrows, nose or mouth. They may require high level accuracy in the location of parts to be ensured on spatial consistency of each part. The number, size and location of the parts to be registered may vary [121][137].

The parts are typically placed as fixed-size blocks around detected landmarks. Optionally, faces may be warped onto a reference frontal face model before patches are

cropped [137][93]. Also, some techniques may apply part detection individually and independently [142].

Among face detection algorithms, Viola and Jones method [128] has gained remarkable attention. The basic idea of that method is the use of a boosted cascade of Haar features which are constructed by considering human face similarities. Another popularly used and recent method is proposed by Zhu et al. [147] in 2012. It is based on mixtures of trees with a shared pool of parts. They modeled every facial landmark as a part and used global mixtures to capture topological changes due to viewpoint.

Point based registration involves the localization of crucial points and has an important role for registration of shape representations. Besides Active Appearance Model, facial feature detectors ([126], [130]) are also widely used methods for point based registration.

2.4.2 Feature Representations

Shape representation is one of the earliest methods that were being used in image processing. In images, shapes can be described by using only shape boundaries and its features (e.g length) or by using the description of the shape region occupied by the object. One of most popular shape representation is chain code representation and it is introduced by Freeman [47] in 1961. Chain code describes an object by a sequence of unit-size line segments with a given orientation. Another contour shape descriptor has been proposed by Peura and Iivarinen [103]. They used ratio of principle axis, circular variance and elliptic variance as descriptors. Some other examples of shape representation can listed as [81], [106] and [61].

Low-level histogram representations are based on extraction of local features of uniform, small regions. In this methodology, the features of each region are represented by local histograms. The final representations are obtained by concatenating all local histograms. Low-level histogram features have some advantages. They are extracted from small regions and robust to global illumination variations to a degree. Local Binary Patters (LBP) [1], Local Phase Quantization (LPQ) [2] and Histogram of Gradients [52] are typically used as Low-Level Histogram representations.

A Gabor function was proposed by Dennis Gabor in 1946 [48] and frequently used for feature extraction, especially in texture-based image analysis and more practically in face recognition [7],[105], [120]. In a typical ***Gabor representation***, an image is filtered

with a set of Gabor filters which have different orientations and spatial frequencies that cover appropriately whole spatial frequency domain. Then the responses from Gabor filters are used for further analysis.

Mainly, ***Bag-of-Words Representation*** (BoW) [72], is a representation used in natural language processing. In this representation, a text can be represented as the bag of its words. Recently, by using same idea, the bag-of-words model has also been used for computer vision. Firstly, several local patches are extracted from images and treated as candidates for basic elements, "words". Then these "words" are converted to "codewords" to produce a "codebook" by using clustering methods. Codewords are defined as the centers of the learned clusters. Each patch in an image is mapped to a certain codeword through the clustering process.

Table 2.3: Popular methods that are being used on facial expression recognition.

Gabor [72][73]
Local Binary Pattern (LBP) [74]
Haar Features [75][76]
Histogram of oriented gradients (HoG) [77]
Discrete Fourier transform [78]
Active appearance model (AAM) [79]
Principle Component Analysis (PCA) [80]
Candid Grid Node [81]
Scale-invariant feature transform (SIFT) [82]
Location of landmark points [83], [84]
Point Distribution Model [85]
Active shape model (ASM) [86]
Optical flow [87]
Active appearance models (AAM) [88]
Shape movements [89]
Facial animation parameters (FAP) [90]

Although most of the representations describe local textures, some approaches aim to obtaining ***data-driven higher level representations***. These approaches try to extract semantically meaningful representations in terms of affect recognition perspective. Non-

negative matrix Factorization (NMF) [94] and sparse coding [27] are two examples of these kind of representations.

Part-based representations process faces in terms of independently registered parts. They ignore the spatial relations among these registered parts. Ignoring the spatial relationships reduces the sensitivity to head-pose variation. Part-based representations proved successful in spontaneous affect recognition tasks (e.g. AU recognition [62], [137] or dimensional affect recognition) where head-pose variation naturally occurs.

2.4.3 Facial Emotion Recognition Studies

There are many studies in the literature for recognizing emotions from face images and videos. It is hard to provide exhaustive coverage of all past efforts in the field of automatic affect recognition. Below we give a short overview of some popular methods.

In 2000, Pantic and Rothkrantz [101] proposed a person independent system that performs recognition and emotional classification from a still full-face image. In the proposed system, dual view face model (frontal-view and side-view) and multiple feature detection techniques are applied in parallel.

Hu et al. [59], in 2002, employed a hybrid approach of NN–HMM. Gabor wavelets were used to extract features from face images and a Multilayer Perceptron (MLP) Neural Network (NN) was used to classify the feature vector into different states of a HMM.

In 2003, Cohen et al. [25] introduced different Bayesian network classifiers by focusing on changes in distribution assumptions, and feature dependency structures. They used Naive–Bayes classifiers and change the distribution from Gaussian to Cauchy, and use Gaussian Tree-Augmented Naive Bayes (TAN) classifiers to learn the dependencies among different facial motion features. Also they proposed a new architecture of hidden Markov models (HMMs) for automatically segmenting and recognizing human facial expression from video sequences.

In 2005, Cowie et al. [28] used FAP (Facial Animation Parameters) for confidence-based feature extraction system and created a fuzzy rule based system for classifying facial expressions.

In 2005, Zhang et al. [144] explored the use of multisensory information fusion technique with Dynamic Bayesian networks (DBNs) for modeling and understanding the temporal behaviors of facial expressions in image sequences. They focused their

attention not only on the nature of the deformation of facial features, but also on features' temporal evolution with human emotions.

In 2006, Yeasin et al. [136] presented a spatiotemporal approach. The proposed approach relies on a two-step strategy on the top of projected facial motion vectors obtained from video sequences of facial expressions. First a linear classification bank was applied on projected optical flow vectors and decisions made by the linear classifiers were coalesced to produce a characteristic signature for each universal facial expression. The signatures thus computed from the training data set were used to train discrete hidden Markov models (HMMs) to learn the underlying model for each facial expression.

In 2009, Kai et al. [20] proposed a graphical model which based on the hidden conditional random fields (HCRFs) where we link the output class label to the underlying emotion of a facial expression sequence, and connect the hidden variables to the image frame-wise action units.

Table 2.4: A list of representative works in the field of video emotion recognition

Researcher(s)	Features	Classification	Database	Acc. (%)
Feng et al. [108]	Face histogram	LP linear programing	Cohn-Kanade	91
Pantic et al. [101]	FACS	Rule based system	Self-defined	91
Hu et al. [59]	Gabor wavelet	HMM + NN	Self-defined	95
Cohen et al. [25]	Facial movements	Tree-Augmented Naive–Bayes	Self-defined Cohn-Kanade	66
Cowie et al. [28]	FACS	Tree Bayesian network	Cohn-Kanade	91
Zhang et al. [144]	FACS	Phase network	100,000 frames	84
Yeasin et al. [136]	Optical flow	HMM	Cohn-Kanade	90
Kai et al. [20]	FACS	HCRF	Cohn-Kanade	93
Senechal et al. [112]	LGBP histogram	SVM	GEMEP-FERA	65
Ulukaya et al. [122]	Coordinate based features (CBF), Distance and angle based features (DABF)	SVM	Cohn-Kanade	88
Chi et al. [58]	Motion tracking	Hough forests	Cohn-Kanade	89

In 2012, Senechal et al. [112] proposed to combine different types of features to automatically detect action units (AUs) in facial images. They used one multikernel support vector machine (SVM) for each AU to detect. They combined spatial-independent feature extraction (LGBP histograms) and statistical spatial shape and texture information (AAM coefficients).

In 2012, Ulukaya et al. [122] presented a Gaussian Mixture Model (GMM) fitting method for estimating the unknown neutral face shape for frontal facial expression recognition using geometrical features. Also in 2014, they proposed a general solution to the baseline problem by estimating the unknown neutral face shape of an expressive face image using a dictionary of neutral face shapes [123].

In 2013, Chi et al. [58] analyzed the non-rigid morphing facial expressions and tried to eliminate the person-specific effects through patch features extracted from facial motion due to different facial expressions. They introduced a 3-D spatial-temporal local feature extraction method for identifying the facial expression by applying Hough forests. They also applied the ROI filtering to reduce the error during the training process and increase the discriminative capacity of the parameter voting.

2.5 AUDIO-VISUAL AFFECT RECOGNITION

The audio-visual emotion recognition methods in the literature have shown the advantages of fusing audio and video modalities [6][29][65][78][133][140]. Below, we review the most commonly used methods for fusing the two modalities.

2.5.1 Multimodal Fusion Methods

2.5.1.1 Feature level fusion

If there are multiple features that are coming from various uncorrelated and independent modalities, feature level fusion of these features sets become meaningful. Mostly it is achieved by concatenating the feature sets. But interior to fusion the combined features must be converted to same format, and same scale.

2.5.1.2 Score level fusion

When scores are consolidated in order to arrive at a final decision, fusion is said to be done at score level. This is also known as measurement level or confidence level fusion. Fusion at score level is the most commonly used approach for combining scores (probabilities) that are coming from more than one classifiers or systems. Although there are many techniques to combine multiple scores, some and popularly used ones can be listed as minimum, maximum, sum, mean and average rule.

2.5.1.2.1 Minimum rule

These techniques simply choose the minimum of the conditional probabilities generated by each of the *i* classifiers while there are total *n* classifiers. The decision for an observed sample x is chosen to be the class ω^* for which has the largest probability between all K probabilities as shown below,

$$P(\omega_k|x) = \min_i\{P(\tilde{\omega}_k|x,\lambda_i)\}, \qquad k = 1,\dots,K \tag{2.1}$$

$$\omega^* = \max_k\{P(\omega_k|x)\}, \qquad k = 1,\dots,K \tag{2.2}$$

where x represents the features of the test data, ω and $\tilde{\omega}$ represent the predicted output labels after and before fusion, $P(\tilde{\omega}_k|x,\lambda_i)$ is the probability of class k for each individual classifier λ_i and ω^* is the final estimated class of the test data.

2.5.1.2.2 Maximum Rule

It is very similar to minimum rule. In this technique the only difference is simply choosing the maximum of the conditional probabilities derived from each i classifier.

$$P(\omega_k|x) = \operatorname{Max}_i\{P(\tilde{\omega}_k|x,\lambda_i)\}, \qquad k = 1,\dots,K \tag{2.3}$$

2.5.1.2.3 Sum Rule

The sum rule sums up the probabilities given to each class in order to generate total probabilities of all classifies.

$$P(\omega_k|x) = \sum_{i=1}^{n} P(\tilde{\omega}_k|x,\lambda_i), \qquad k = 1,\dots,K \tag{2.4}$$

2.5.1.2.4 Mean Rule

By averaging the probabilities given to each class, we obtain the mean rule. $\frac{1}{n}$ serves as normalization factor.

$$P(\omega_k|x) = \frac{1}{n}\sum_{i=1}^{n} P(\tilde{\omega}_k|x,\lambda_i), \qquad k = 1,\dots,K \tag{2.5}$$

2.5.1.2.5 Weighted Average Rule

By additionally adding a classifier weights $P(\lambda_i|x)$, we would have the weighted average method.

$$P(\omega_k|x) = \frac{1}{n}\sum_{i=1}^{n}[P(\widetilde{\omega}_k|x,\lambda_i)\,P(\lambda_i|x)] \quad , \qquad k = 1,\dots,K \tag{2.6}$$

2.5.1.2.6 Product Rule

Although it is similar to sum rule, this time the probabilities given to each class by all classifiers are multiplied to generate final probability of one class.

$$P(\omega_k|x) = \prod_{i=1}^{n} P(\widetilde{\omega}_k|x,\lambda_i) \quad , \qquad k = 1,\dots,K \tag{2.7}$$

2.5.1.2.7 Bayesian Framework

Bayes' theorem says that joint probabilities are the products of conditional and marginal probabilities. And in bimodal recognition system, each classifier gives conditional probability for each class. These conditional probabilities can be combined based on a Bayesian weighting framework and the results for each class may be used as a final decision of the emotional state of the each sequence.

$$P(\omega_k|x) = \sum_{i=1}^{n}\sum_{k=1}^{K}[P(\omega_k|\widetilde{\omega}_k,\lambda_i)\,P(\widetilde{\omega}_k|x,\lambda_i)\,P(\lambda_i|x)] \quad , \qquad k = 1,\dots,K \tag{2.8}$$

2.5.2 Multimodal Emotion Recognition Studies

Theoretically and empirically, many studies have demonstrated the advantage of the integration of multiple modalities, like vocal and visual expressions, in human affect perception over single modalities [4], [107]. Below, we give a brief overview of the approaches used for multimodal emotion recognition.

Chen et al. [23] proposed a method in 1998 for emotion recognition from audio-visual signal. They used 16 prosodic audio features and followed the horizontal and vertical positions of the eye brow, check lifting and size of the mouth opening.

In 2000, Silva et al. [115] classified six basic emotions by combining audio and video signals and using a rule based system. They observed 72% emotion recognition accuracy for multimodal audio-visual recognition, while audio and video recognition accuracies were 32% and 62% respectively.

In 2002, Lisetti and Nasoz [76] combined facial expression and physiological signals to recognize the user's emotions, like fear and anger, and then adapted an animated interface agent to mirror the user's emotion.

Duric et al. [40] applied a model of embodied cognition that can be seen as a detailed mapping between the user's affective states and the types of interface adaptations.

In 2004, Busso et al. [16] worked on feature and decision level integration of speech and facial modalities. In their studies, fusion of the audio and visual data improved the performance of visual only system 5 percent.

Chen et al. [22] fused audio and visual features at feature level by using two ways. Initially, they combined the features by concatenating the audio and visual information. And as a second method, they made the size of feature vectors of both modalities same by duplicating the audio features since audio features' amount was the half of visual features.

Paleari and Lisetti [100], in 2006, presented a framework for multimodal emotion recognition that could accept new recognition modules based on Scherer theories.

Schuller and Wimmer [109] proposed an audio-visual emotion recognition system that uses feature space combination. In the proposed system, audio and video features are firstly derived as Low-Level-Descriptors. Synchronization and feature combination is examined by multivariate time-series analysis.

The proactive HCI tool of Maat and Pantic [84] is capable of learning and analyzing the user's context-dependent behavioral patterns from multisensory data and adapting the interaction accordingly

Kapoor et al. [67] combined information from cameras, a sensing chair, and mouse and wireless skin sensor to detect frustration when the user needs help.

In 2010, Mansurizadeh et al. [86] proposed an asynchronous feature level fusion approach that creates a unified hybrid feature space out of the individual signal measurements. They tested their approach on EMODB and eNTERFACE'05 databases and achieved 82% and 84% emotion recognition accuracies, respectively.

Gajsek et al. [49] proposed an audio-visual emotion recognition system in 2010. They used prosodic, spectral and cepstrum features as audio features. They proposed a video subsystem that does not rely on the tracking of specific facial landmarks. Gajsek used

SVM classifier and tested on eNTERFACE database which resulted 71.3% emotion recognition accuracy.

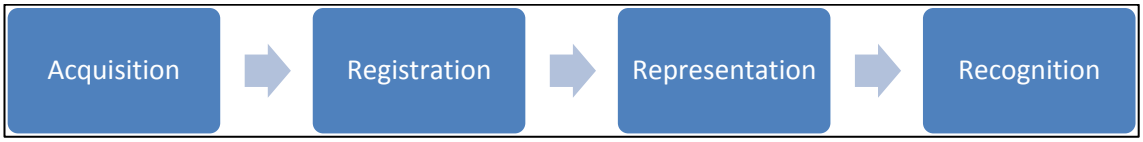
In 2014, Zhalehpour et al. [140][141] proposed a framework for multimodal emotion recognition based on automatic peak frame selection from audio-visual sequences. They evaluated the performance of their approach on eNTERFACE'05 and BAUM-1 databases and got 76.4% and 64.05% recognition accuracies, respectively.

3. AFFECT RECOGNITION USING KEY FRAME SELECTION BASED ON MINIMUM SPARSE RECONSTRUCTION

A visual affect recognition system can be broken down into some fundamental components as acquisition, registration, representation and recognition. All procedures start with an acquisition of a video or an image. A video is a collection of multiple consecutive frames that follow each other with small time changes. So they can be handled as a frame sequence.

Most of the time, frames contain more than one objects, background scenes and even unintended human bodies and faces other than target face. Besides detecting the target face in each frame, there is a need for registration to align the detected faces to minimize translation, scale and head-pose differences. So the registration step is a fundamental step of an audio-visual affect recognition system. Some different face registration methods that are being used in literature were mentioned in Section 2.4.1. The second step is the representation of the face region to analyze the facial expression. Several popular methods for face representation were mentioned in Section 2.4.2. The recognition step produces the final typical output of an affect recognition system, which is the label of the facial action or an emotion. This is performed by a machine learning technique. A general framework for audio-visual affect recognition is described in Figure 3.1.

Figure 3.1: General emotion recognition framework



In this chapter, the details of the proposed affect recognition method using key frame selection based on minimum sparse reconstruction will be given. In the following, each block in Figure 3.1 will be explained in detail.

3.1 PREPROCESSING OF FACE IMAGES

Since it is so common to have head motion between the frames of a facial expression sequence, there is a need for registration by aligning the faces among subsequent

frames. Detection and cropping of the face region are also helpful for elimination of unnecessary regions such as the background and hair.

One of the methods that can be used for basic face alignment is based on landmark locations detection on the face as seen in

Figure 3.2. By using the centers of the eyes, translation and rotation differences between detected faces can be eliminated. Similarly, by fixing the distance between two eye centers to some specific distance scaling differences also can be eliminated.

Figure 3.2: Cropping and resizing of an image, based on eye center locations

In our framework, we used Zhu's [147] face tracker method to detect the landmarks. This method gives 68 landmarks for frontal faces and the eyes are represented using six landmarks. The centers of the eyes are calculated by averaging these six eye landmark points around the eye. In order to compensate for any scale differences between the frames, images are scaled to obtain an inter-ocular distance of 64 pixels. Then, images are aligned based on eye locations and cropped in a way such that the face region has a size of 168×126 (see

Figure 3.2).

3.2 EXTRACTION OF VISUAL FEATURES FROM FACE IMAGES

3.2.1 Local Phase Quantization Features

Local Phase Quantization (LPQ) operator was proposed by Ojansivu and Heikkila [97] in 2008. It was originally proposed for blur insensitive texture classification but it has been shown that it is also useful for emotion recognition [33][135][29]. In this thesis, the LPQ method is used to construct a face descriptor representing the feature of the face.

Let us assume that we are given a (degraded) image, on which we want to calculate a texture descriptor. In digital image processing, an observed blurred image can be represented as the convolution [8] of the original image and point spread function of the blur.

$$g(\boldsymbol{x}) = (f * h)(\boldsymbol{x}) \tag{3.1}$$

where $f(\boldsymbol{x})$ is the original image, $g(\boldsymbol{x})$ is the observed image and $h(\boldsymbol{x})$ is the point spread function (PSF) of the blur.

In the frequency domain, this corresponds to

$$G(\boldsymbol{u}) = F(\boldsymbol{u}).H(\boldsymbol{u}) \tag{3.2}$$

where $F(\boldsymbol{u})$, $G(\boldsymbol{u})$ and $H(\boldsymbol{u})$ are the discrete Fourier transforms (DFT) of the original image $f(\boldsymbol{x})$, the blurred image $g(\boldsymbol{x})$, and the point spread function $h(\boldsymbol{x})$ of the blur.

If $h(\boldsymbol{x})$ is centrally symmetric, which means $h(\boldsymbol{x}) = h(-\boldsymbol{x})$, Fourier transform of $h(\boldsymbol{x})$ is always real valued and phase angle is only two valued function as given by

$$\angle H(\boldsymbol{u}) = \begin{cases} 0 \; if \; H(\boldsymbol{u}) \geq 0 \\ \pi \; if \; H(\boldsymbol{u}) < 0 \end{cases} \tag{3.3}$$

The local phase quantization (LPQ) method is based on the blur invariance property of the Fourier phase spectrum. LPQ uses the local phase information extracted using the 2D Fourier transform computed over an M-by-M neighborhood N_x at each pixel position *x*. So the computed Fourier transform is defined by

$$F(\boldsymbol{u}, \boldsymbol{x}) = \sum_{\boldsymbol{y} \in N_x} f(\boldsymbol{x} - \boldsymbol{y})\, e^{-j2\pi \boldsymbol{u}^T \boldsymbol{y}} = \boldsymbol{w}_{\boldsymbol{u}}^T \boldsymbol{f}_{\boldsymbol{x}} \tag{3.4}$$

where $\boldsymbol{w}_{\boldsymbol{u}}$ is the basis vector at frequency $\boldsymbol{u}$ and $\boldsymbol{f}_{\boldsymbol{x}}$ denotes the vector containing the values of all M^2 image samples, which come from N_x.

In LPQ feature extraction only four complex coefficients are considered, corresponding to 2D frequencies $\boldsymbol{u}_1 = [a, 0]^T$, $\boldsymbol{u}_2 = [0, a]^T, \boldsymbol{u}_3 = [a, a]^T, and \; \boldsymbol{u}_4 = [a, -a]^T$, where *a* is a small frequency at which $H(\boldsymbol{u}) \geq 0$.

Let

$$\boldsymbol{F}'_{\boldsymbol{x}} = [F(\boldsymbol{u}_1, \boldsymbol{x}), F(\boldsymbol{u}_2, \boldsymbol{x}), F(\boldsymbol{u}_3, \boldsymbol{x}), F(\boldsymbol{u}_4, \boldsymbol{x})] \tag{3.5}$$

$$\mathbf{F}_{\boldsymbol{x}} = [\mathrm{Re}\{\boldsymbol{F}'_{\boldsymbol{x}}\}, \mathrm{Im}\{\boldsymbol{F}'_{\boldsymbol{x}}\}]^T \tag{3.6}$$

So that,

$$\boldsymbol{W} = \left[\mathrm{Re}\{\boldsymbol{w}_{u_1}, \boldsymbol{w}_{u_2}, \boldsymbol{w}_{u_3}, \boldsymbol{w}_{u_4}\}, \mathrm{Im}\{\boldsymbol{w}_{u_1}, \boldsymbol{w}_{u_2}, \boldsymbol{w}_{u_3}, \boldsymbol{w}_{u_4}\}\right]^T \tag{3.7}$$

$$\mathbf{F}_{\boldsymbol{x}} = \boldsymbol{W}_{\boldsymbol{u}} \boldsymbol{f}_{\boldsymbol{x}} \tag{3.8}$$

$$\mathbf{G}_{\mathbf{X}} = \mathbf{V}^T \mathbf{F}_{\mathbf{x}} \tag{3.9}$$

Where **V** is an orthonormal matrix derived from the singular value decomposition (SVD).

Next, $\mathbf{G}_{\mathbf{x}}$ is computed for all image positions, i.e., $\mathbf{x} \in \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N\}$, and the resulting vectors are quantized using a simple scalar quantizer;

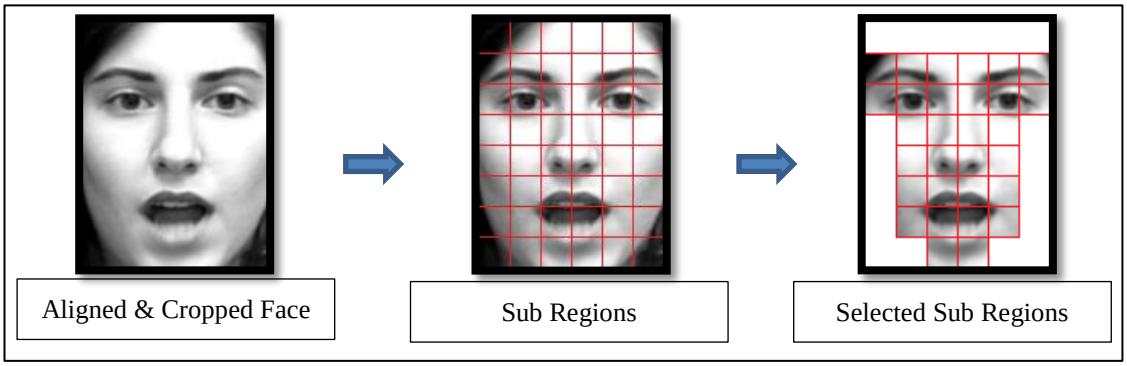
$$q_i(\boldsymbol{x}) = \begin{cases} 1, if\ g_i(\boldsymbol{x}) > 0 \\ 0, \quad otherwise \end{cases} \tag{3.10}$$

where, $g_i(\boldsymbol{x})$ is the jth component of $\mathbf{G}_{\mathbf{x}}$. The quantized coefficients are represented as integer values between 0-255 using binary coding as follows:

$$f_{LPQ}(\boldsymbol{x}) = \sum_{i=1}^{8} q_i(\boldsymbol{x}) 2^{i-1} \tag{3.11}$$

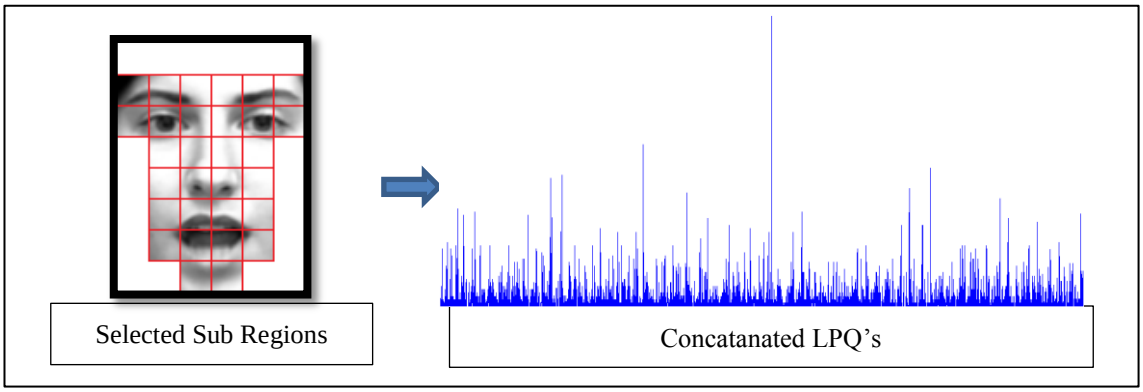
Traditionally, face can be split into several sections such as forehead, cheeks, eyes, nose and chin. However some parts of face do not change very much with emotions and do not carry information about the facial expression as upper forehead, outer sections of cheek and jaw. In order to separate emotion related regions more accurately, we divide the face image into 48 sub-blocks of size 8×6 and select 30 sub-blocks that are more relevant with emotion representation as shown in Figure 3.3.

Figure 3.3: Sub region selection for feature extraction



In the selected emotion related sub regions, we extract the 256 bin histogram of LPQ features of each region. The LPQ features of the 30 sub-blocks are concatenated into a long vector of histogram to form a single feature vector shown in Figure 3.4. The length of the final histogram of the whole image is 7680 (256 bin histogram x 30 sub-blocks) and it is used as the facial expression feature. We used the implementation available from [150] and used it with default parameters (window size is 3 × 3, DFT is calculated using a uniform window, a = 0.7).

Figure 3.4: LPQ feature extraction from selected sub regions



3.2.2 LBP TOP Features

Local binary patterns (LBP) is a feature extraction method used for classification (texture etc.) purposes in computer vision that was proposed in 1990 [36] [96]. Recently, it has been applied to face recognition [3] and facial expression recognition [1] [113]. While the original LBP was only designed for static images, LBP-TOP (Local

Binary Pattern histograms from Three Orthogonal Planes) has been used for dynamic textures and facial expression recognition [145].

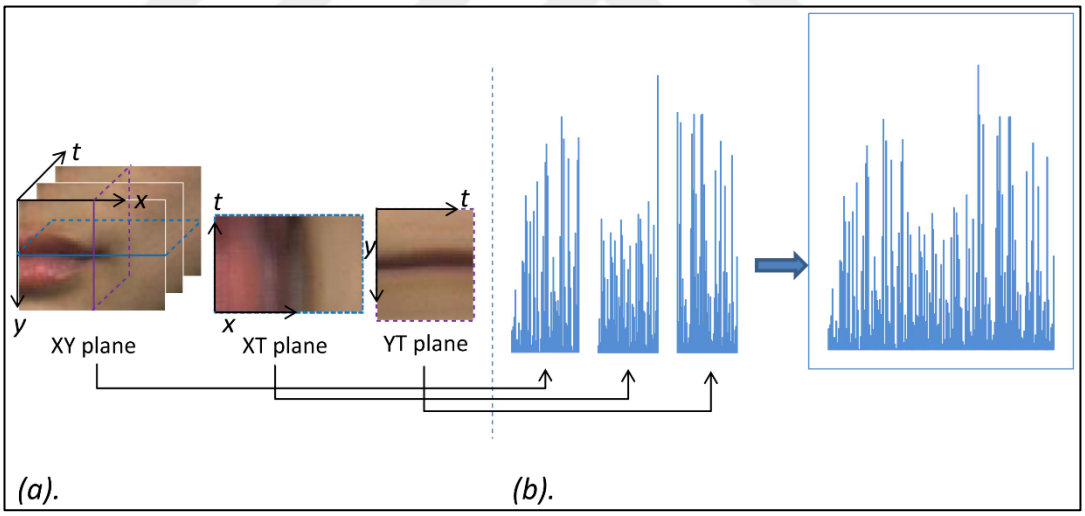
Given a pixel located at $\boldsymbol{x}$, its LBP code is computed as:

$$LBP(\boldsymbol{x}) = \sum_{p=0}^{P-1} s(i_{\boldsymbol{p}} - i_{\boldsymbol{x}})2^P \tag{3.12}$$

$$s(x) = \begin{cases} 1 \; if \; x \geq 0 \\ 0 \; if \; x < 0 \end{cases} \tag{3.13}$$

where $i_{\boldsymbol{x}}$ denotes the intensity of the center pixel, P denotes the total number of neighbors of $\boldsymbol{x}$ parameterized by the radius of the neighborhood R, while $i_{\boldsymbol{p}}$ denotes the intensity of the neighbouring pixels. Then, the histogram of all LBP patterns is computed for all pixels in an image.

Figure 3.5: The texture of LBP-TOP planes and the corresponding histograms



(a) XY, XT and YT planes of a micro-expression sample
(b) concatenated LBP-TOP feature

Source: Yan, Wen-Jing; Li, Xiaobai; Wang, Su-Jing; Zhao, Guoying; Liu, Yong-Jin; Chen, Yu-Hsin; Fu, Xiaolan (2014): The texture of three planes and the corresponding histograms.

LBP-TOP computes the local spatio-temporal patterns based on LBP. LBP-TOP feature is constructed by the concatenation of LBP histograms on three orthogonal XY, XT and YT planes. The XT and YT planes contain the temporal transition information pertaining to the facial movement displacement e.g. how eyes, lips, muscles or

eyebrows change over time. In contrast, the XY plane contains only spatial information which includes both expression and identity information of a face appearance (see Figure 3.5).

3.3 EXTRACTION OF AUDIO FEATURES

3.3.1 Mel Frequency Cepstral Coefficients (MFCC) Features

Researches showed that human audio perception sensitivity is not linear [12]. Humans are more sensitive to low frequency components than high frequency components. In order to analyze the whole spectrum in sufficient detail, Mel Frequency Cepstral Coefficients (MFCCs) handle the frequency spectrum on a nonlinear mel scale of frequencies which is based on the human ear scale.

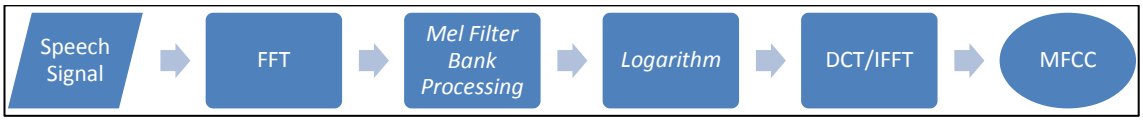
In MFCC computation, the speech signal is initially divided into time frames consisting of an arbitrary number of samples (e.g. 25msec frames). Generally, overlapping windows are used to smooth transitions from one window to another. Then, each windowed frame is smoothed with a Hamming window to eliminate discontinuities at the window edges.

After the windowing, Fast Fourier Transformation (FFT) is calculated for each window to extract the frequency components. The logarithmic Mel-Scaled filter bank is applied to the Fourier transform. This scale is approximately linear up to 1 kHz, and logarithmic at greater frequencies. The relation between frequency of speech and Mel scale can be established as:

$$\text{Mel Scaled Frequency} = 2595 \log (1+f\,(\text{Hz})/700 \tag{3.14}$$

The last step is to calculate the Discrete Cosine Transformation (DCT) of the outputs from the filter bank. DCT sorts coefficients according to their significance and the first coefficient is excluded from MFCCs because of its unreliability. The overall procedure of MFCC extraction is shown on Figure 3.6.

Figure 3.6: The structure of MFCCs



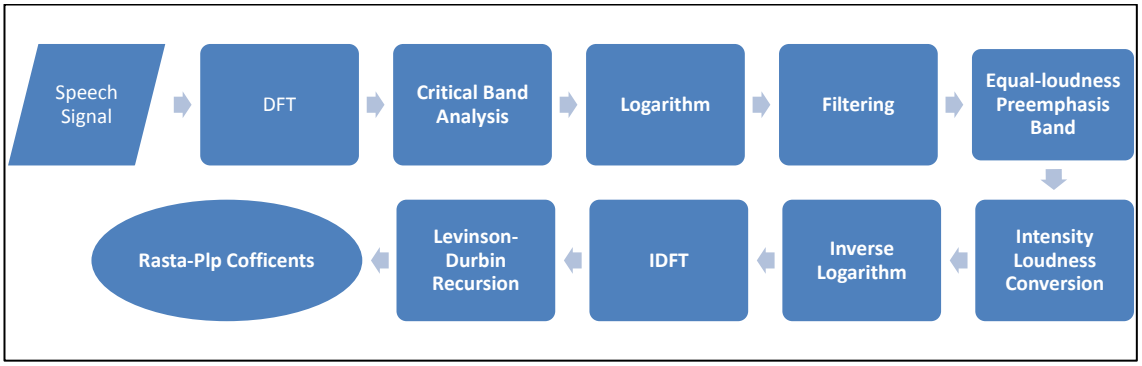
In our framework, we applied a Mel filter with order 12 and used generated 12 MFCC features. In addition to these MFCCs, first and second derivatives of the features are also calculated to capture local dynamics. Then, nine statistical functions, namely, maximum, minimum, maximum position, minimum position, mean, variance, range, kurtosis and skewness are applied to the first 12 MFCCs and their first and second derivatives to extract the MFCC feature vector for an audio segment (e.g. a sentence). All these steps generate a feature vector of length 324 (12x3x9) related to the MFCC.

3.3.2 Relative Spectral Transform – Perceptual Linear Prediction (RASTA - PLP) Features

Relative Spectral Transform (RASTA) [55] is another popular technique which applies a band-pass filter to the energy around frequencies for suppressing short-term noise. One of the popular speech features used in emotion recognition is known as RASTA-PLP (Relative Spectral Transform - Perceptual Linear Prediction). It is a combination of PLP and RASTA methods.

PLP (Perceptual Linear Predictive) was proposed by Hermansky [56] and designed for minimizing the differences between speakers while preserving the main speech information. The steps for extraction of RASTA-PLP features for each frame are shown in Figure 3.7.

Figure 3.7: The structure of RASTA-PLP Method

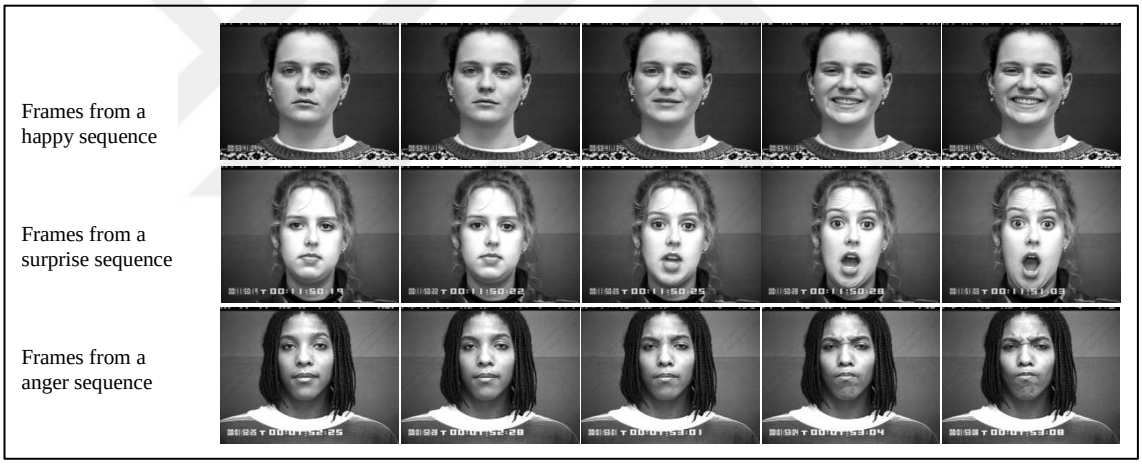


In our framework, the first 13 RASTA-PLP coefficients are calculated by using filter of order 20. Similar to MFCC features, their delta and double delta features are appended, as well. Also the same statistical parameters as used for MFCCs are calculated for the RASTA-PLP coefficients. These steps create 351 (13x3x9) RASTA-PLP related features. Then the MFCC and RASTA-PLP related feature vectors are concatenated in order to have audio feature vectors of length 675 (324 + 351).

3.4 KEY FRAME SELECTION BASED ON VIDEO SUMMARIZATION

An audio-visual video with an emotional expression consists of many frames, where each frame may represent different emotions with different intensities. Moreover, given a video that contains a single emotional expression, some frames may represent the emotion with a high intensity while others may be close to neutral (see Figure 3.8).

Figure 3.8: Several sequences from CK+ database containing single emotion

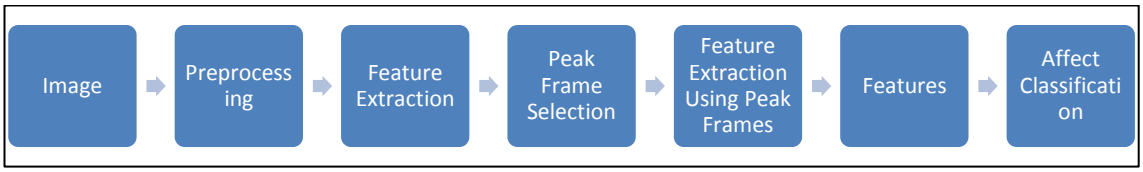


The frame rate of a video may affect the recognition accuracy of expression movements [104] and small expression changes between consecutive frames may influence affect recognition performance. In order to shorten the necessary time to process the video and also avoid some false affect recognition decisions, video summarization methods can be used to represent the whole video using a representative subset of samples.

Video summarization methods may aim to extract only one key frame which can be considered as the most representative one among all others. Other methods may aim to get more than one frames until enough detail is covered by extracted summary frames. One way of reducing the number of frames is down-sampling the frame sequence.

Although this reduced frame sequence can be easily handled by uniform down-sampling, creating a "key frame" sequence and classifying the facial expression on the selected "key frames" may give better results. Therefore, our motivation is summarize the content of the video in the best possible way with as few frames as possible.

Figure 3.9: General framework of peak frame selected affect recognition system



The key frames are selected so that they summarize the content of the video in the best way. This is based on the assumption that there is a single emotion in the sequence, which is true for most databases in the literature. The assumption behind the utilized key frame selection method is that set of key frames are the most representative frames in the sequence among others and also the conceptual information of the sequence is mostly covered by key frames [89]. Therefore, we view the key frame selection as a similar problem to video summarization. In [89], Mei et al. formulated video summarization as a problem of selecting the minimum number of frames to reconstruct the entire video as accurately as possible. We applied this video summarization method to emotional videos to get emotional "key frames".

Given a video with n frames, each frame is a candidate to be a key frame. Let $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n] \in R^{d \times n}$, where $\mathbf{f}_i \in \mathrm{R}^d$ denotes the feature vector corresponding to frame i. The goal of key frame selection is to select an optimal subset $\boldsymbol{F_K} = [\mathbf{f}_{k_1}, \mathbf{f}_{k_2}, \dots, \mathbf{f}_{k_m}] \in R^{d \times m}$ such that $k_1, k_2, \dots, k_m \in [1,2, \dots, n]$. There are two goals when the subset is formed: i) The original video is reconstructed accurately and ii) the number of key frames is as small as possible. That is, the following minimum sparse reconstruction (MSR) expression is minimized:

$$\min_{\mathbf{S}} \frac{1}{2} \|\mathbf{F} - \mathbf{F}_K \mathbf{A}\|_2 + \lambda \|\mathbf{S}\|_0 \tag{3.15}$$

such that $\mathbf{F}_K = \mathbf{FS}$ and $\mathbf{A} = \mathrm{f}(\mathbf{F}, \mathbf{F}_K)$ where $\mathbf{S}$ is a diagonal selection matrix that models the selection of key frames from the original video:

$$S_{ij} = \begin{cases} 0, & i \neq j \\ 0 \text{ or } 1, & i = j \end{cases} \tag{3.16}$$

and $\|S\|_0$ is the L_0 norm of the selection matrix, which is the number of nonzero elements indicating the number of key frames selected. Therefore, scarcity is ensured by the L_0 norm. In (1), **A** represents the reconstruction coefficients of **F** by the matrix $\mathbf{F}_K$ which are computed using the reconstruction function $\mathbf{f}(.,.)$, $\|.\|_2$ represents the L_2 norm, and λ is a weighting coefficient. The first term in (1) tries to minimize the least-square reconstruction error (LSRE), while the second term minimizes the number of key frames selected.

Assuming that m keyframes have been selected, the next keyframe chosen should maximally decrease LSRE. Therefore, the frame which gives the maximum LSRE at the current iteration should be selected as the next key frame:

$$\mathbf{f}_{k_{m+1}} = \arg \max_{\mathbf{f}_j \in \mathbf{F}/\mathbf{F}_K} \left\| \mathbf{f}_j - \mathbf{F}_K \mathbf{a}_j \right\|_2 \tag{3.17}$$

where $\mathbf{a}_j$ represents the reconstruction coefficient for the j^{th} frame and $\mathbf{F}/\mathbf{F}_K$ represents the set of all non-keyframes. This is equivalent to selecting the worst reconstructed frame, after normalization by the vector magnitude:

$$\mathbf{f}_{k_{m+1}} = \arg \min_{\mathbf{f}_j \in \mathbf{F}/\mathbf{F}_K} \frac{\left\| \mathbf{F}_K \mathbf{a}_j \right\|_2}{\left\| \mathbf{f}_j \right\|_2} \tag{3.18}$$

In order to determine the reconstruction coefficients, the orthogonal subspace projection (OSP) method is used [89] by projecting all frames to the space spanned by $\mathbf{F}_K$, which gives:

$$\mathbf{a}_j = (\mathbf{F}_K^T \mathbf{F}_K)^{-1} \mathbf{F}_K^T \mathbf{f}_j = \mathbf{P}_K \mathbf{f}_j \tag{3.19}$$

The algorithm continues to select new key frames as long as the percentage of reconstruction error (POR) of any frame is below a predetermined threshold, i.e.:

$$\text{POR}_j = \frac{\left\| \mathbf{F}_K \mathbf{a}_j \right\|_2}{\left\| \mathbf{f}_j \right\|_2} < T_P \tag{3.20}$$

The overall algorithm that we use for key frame selection can be summarized as follows:

Algorithm: Minimum sparse reconstruction based key frame selection algorithm.

Input: The expressive video $\mathbf{F} \in \boldsymbol{R}^{\boldsymbol{d}\times\boldsymbol{n}}$ with n frames, where each frame is represented by the LPQ features extracted from the face region.

Output: The key frame set $\mathbf{F}_K \in \boldsymbol{R}^{\boldsymbol{d}\times\boldsymbol{p}}$.

1. Initialize the key frame set using the first frame of the sequence as $\mathbf{F}_K = [\boldsymbol{f}_{\boldsymbol{k}_1}]$ and set $m = 1$.
2. Calculate the POR for all frames in set $\mathbf{F}/\mathbf{F}_K$ using (3.20).
3. Repeat steps 4-6 while POR of any frame is smaller than T_P.
4. Select the next key frame using (3.18).
5. Increase m by one.
6. Calculate the POR for all frames in set $\mathbf{F}/\mathbf{F}_K$ using (3.20).

After the iterations terminate, we discard the first frame selected at the initialization step. Some examples from the EmotiW 2015 AFEW 5.0 and CK+ databases for the key frame selection results are shown in Figure 3.10 and Figure 3.11. We can observe that the minimum numbers of frames that represent the whole video have been selected. In the conducted tests, T_P is intuitively selected as 0.8.

Figure 3.10: An example of video summarization from AFEW 5.0 database.

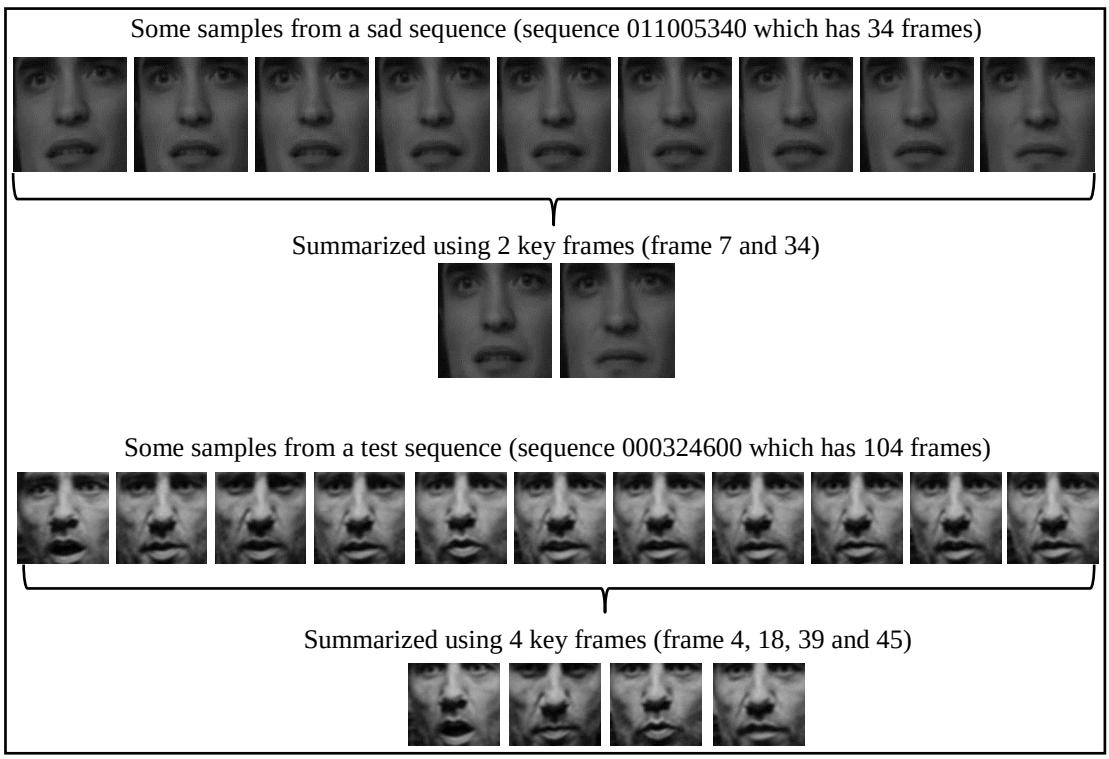
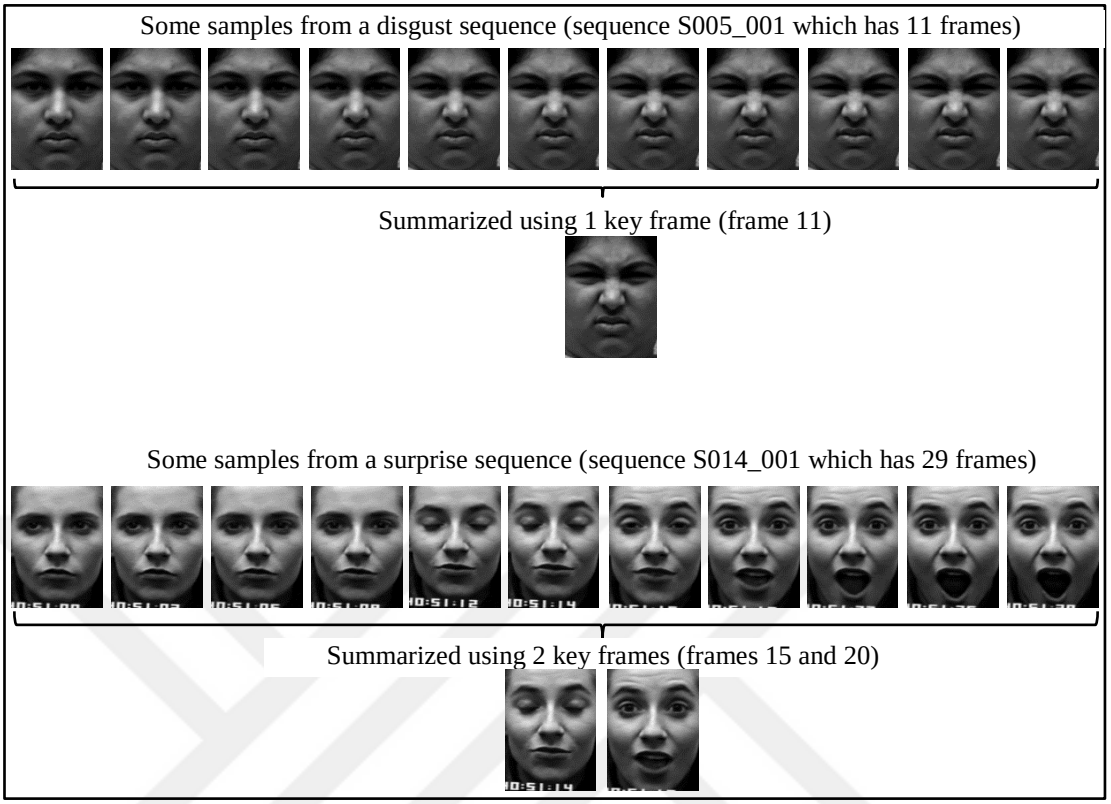


Figure 3.11: An example of video summarization from CK+ database.



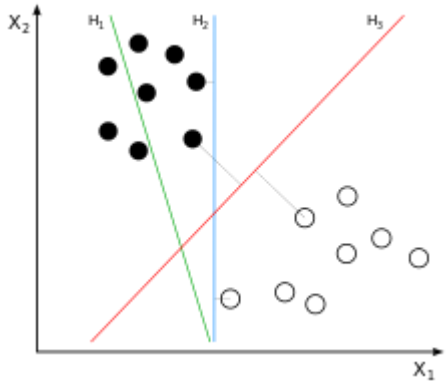
3.5 CLASSIFICATION

In recent years, kernel-based techniques such as support vector machines (SVMs) which are a group of supervised learning methods that can be applied to classification or regression [27] has become very popular among machine learning algorithms.

In a common classification problem, some data points which belong to one of two classes are given and then the aim is to decide the class of a new data point. In the case of SVM, a data point is viewed as a p dimensional vector. And the classifier tries to separate such points with a (p-1) dimensional hyper plane. This is called a linear classifier.

As it is seen in Figure 3.12, there are many hyper planes that might separate the data. One reasonable choice as the best hyper plane is the one that represents the largest separation between the two classes. So the hyper plane is chosen such that the distance from it to the nearest data point on each class is maximized.

Figure 3.12: Binary linear classifier.



H1 does not separate the classes.
H2 does, but only with a small margin.
H3 separates them with the maximum margin.

Source: https://en.wikipedia.org/wiki/Support_vector_machine

In this thesis, we used an SVM classifier implemented in the LIBSVM toolbox [19] to classify each of the audio and video features. In order to classify the audio features, we used an SVM classifier with a radial basis kernel function and one-against all method. Before classification, we normalized the numerical values of audio features to the interval [0, 1] to prevent features with large numeric values dominate features with small numeric values during classification. For the classification of the video features, we used an SVM classifier with a linear kernel to avoid the curse of dimensionality problem, since the dimension of the features is high (i.e. 7680).

In the Emotion Recognition in the Wild (EmotiW 2015) Challenge, we used a score level fusion technique, where we combine the probabilities for each class, which are estimated using each modality separately. We tested several approaches for combining the probabilities [6] estimated using the SVM classifiers and the best results were obtained using the product rule [133], in which the probabilities obtained from the classification of each modality is multiplied for a given test vector and we predict the final label as the one which gives the maximum product:

$$P(\omega_k|x) = \prod_{i=1}^{2} P(\tilde{\omega}_k|x,\lambda_i)\ , \qquad k = 1,\dots,6$$

$$\omega^* = \max_k\{P(\omega_k|x)\}\,, \qquad k = 1,\dots,6$$

where x represents the features of the test data, ω and $\tilde{\omega}$ represent the predicted output labels after and before fusion, $P(\tilde{\omega}_k|x,\lambda_i)$ is the probability of class k for each individual classifier λ_i and ω^* is the final estimated class of the test data (see Figure 3.13 and Figure 3.14).

Figure 3.13: Fusion of LBP-TOP and LPQ video features

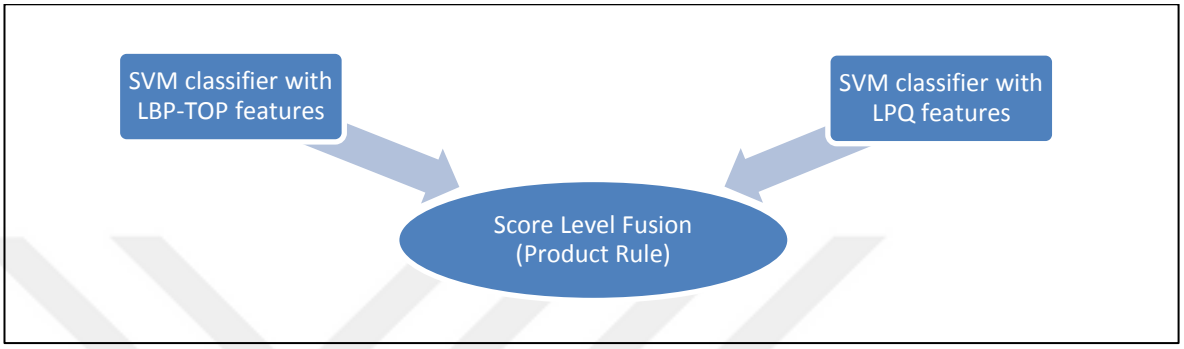
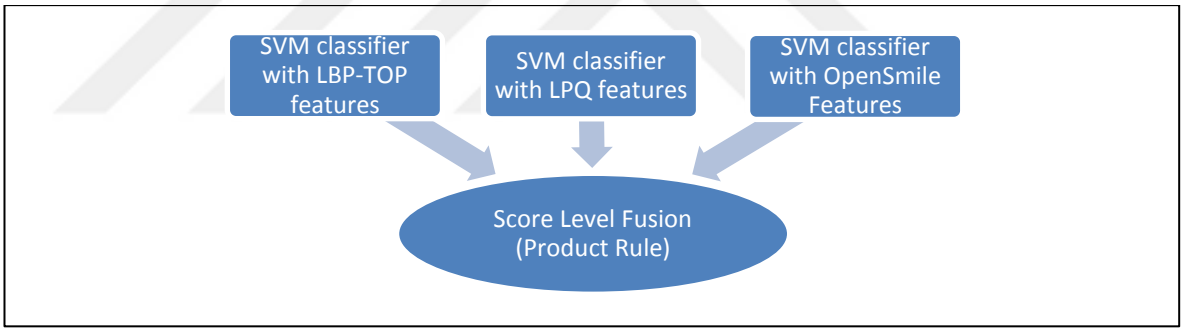


Figure 3.14: Fusion of LBP-TOP, LPQ and OpenSmile audio features



4. EXPERIMENTAL RESULTS

4.1 DATABASES USED

In this chapter, we describe the four databases (CK+, eNTERFACE, BAUM1, AFEW) used in the experiments in more detail and give our key frame selection based affect recognition results on these databases.

4.2 EXPERIMENTAL SETUPS AND RESULTS

4.2.1 Results on CK+ Database

Extended Cohn-Kanade (CK+) [66] dataset was released by the Affect Analysis Group at the University of Pittsburg in 2010. CK+ database includes both posed and non-posed (spontaneous) expressions and additional types of metadata. The posed expressions set contains a total of fully FACS coded 593 frame sequences which belong to 123 subjects. The image sequences vary in duration (between 6 to 71 frames) and start from the neutral facial expression and end at the apex (peak) phase of the expression (see Figure 4.1). The sequences were recorded using a Panasonic WV3230 camera and ages of the subjects are between 18 and 50 years. In the CK+ database, 327 of 593 sequences have been labeled with one of the seven discrete emotions: anger, contempt, disgust, fear, happiness, sadness and surprise.

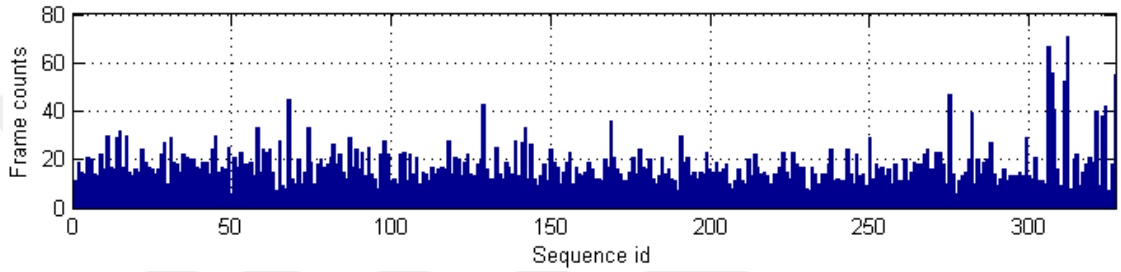
Figure 4.1: Expression intensity among frame sequences of CK+ Database

Source: http://www.pitt.edu/~emotion/ck-spread.htm

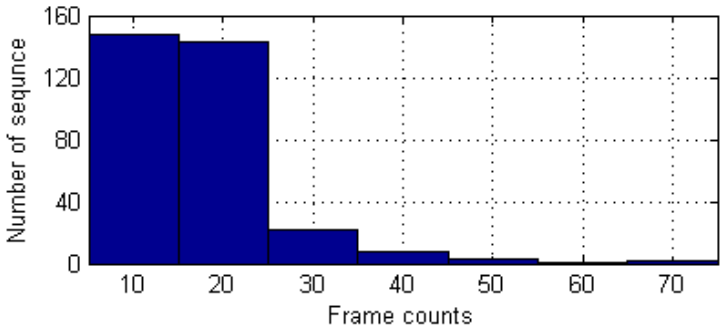
The CK+ image sequences are annotated with 68 landmark points. Some examples for the landmarks that are given in the CK+ database can be seen in Figure 4.2. The distribution of the number of frames for each sequence in CK database can be seen in Figure 4.3.

Figure 4.2: The 68 landmarks given in the CK+ database

Figure 4.3: Number of frames in CK+ sequences



(a) Number of frames in each sequence



(b) Distribution of the number of frames

The minimum and maximum number of frames in a sequence are 6 and 71, respectively. The average number of frames in the sequences is 18 (Table 4.1).

Table 4.1: Number of frames for CK+ sequences

Min	Max	Mean
6	71	18

The tested databases have many frames in each sequence. We tested the key frame selection method on the CK+ database to summarize each sequence. The summary of the key frame extraction algorithm was given in Section 3.3. In the applied video

summarization method, it is not guaranteed to have only one frame the summary which is the key frame whose expressed emotion impression is at the highest level in the sequence. The aim is to select the minimum number of frames that represent the whole sequence in the best possible way. Two examples of key frame selection can be seen in Figure 4.4 and Figure 4.5 for the CK+ database.

Figure 4.4: Key frame extraction when two frames are selected.

(Surprised of subject 77 in CK+ database)

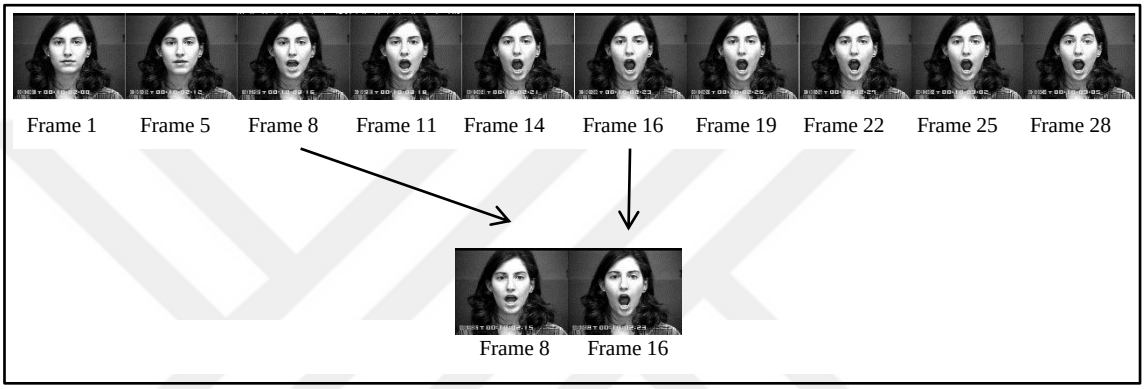
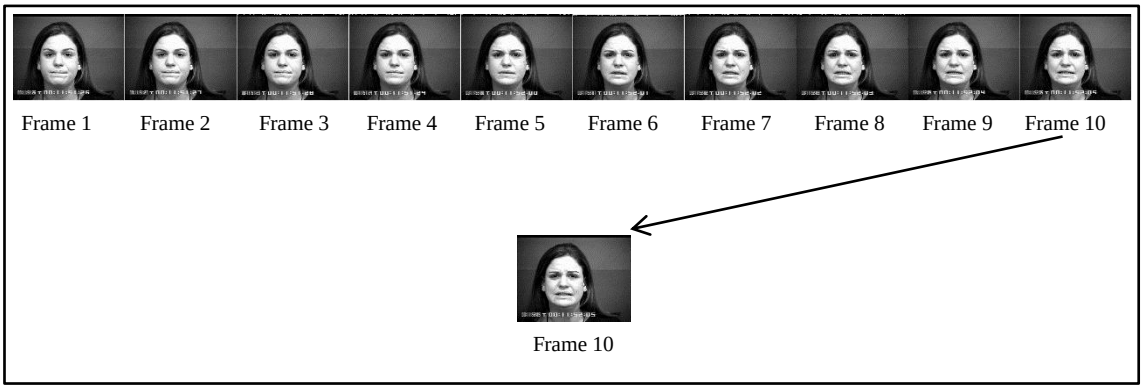


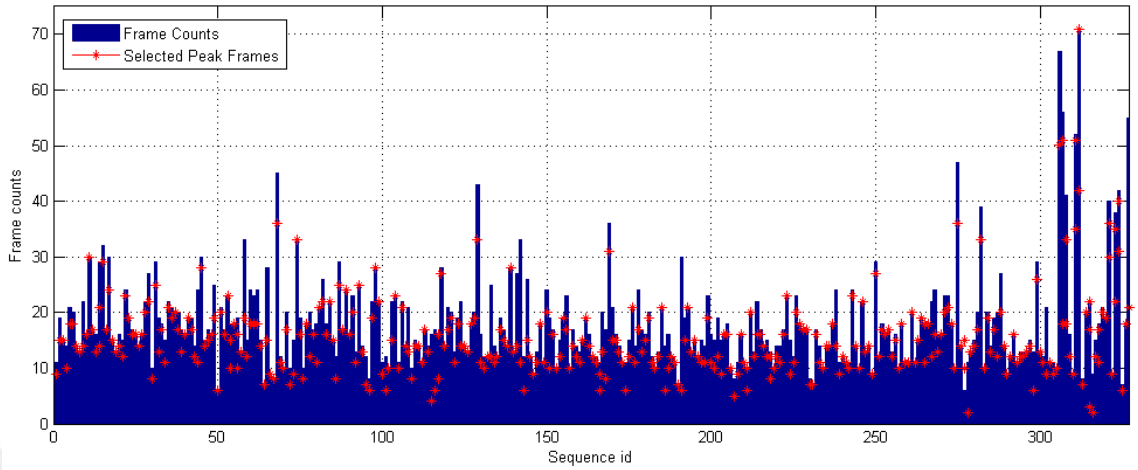
Figure 4.5: Key frame extraction when only one (last) frame is selected.

(Fear of subject 68 in CK+ database)



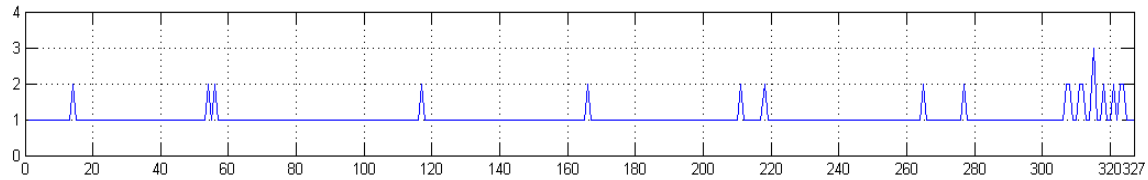
In CK+ database, it is known that each of the sequences contains images from onset (neutral frame) to peak expression (last frame). So it is inherently expected to have the last frame as key frame as a result of key frame extraction algorithm. Unfortunately, by using the video summarization method in [89], the last frame is not always selected as key frame. The selected key frames are marked on the frame distribution in Figure 4.6.

Figure 4.6: Selected key frames are marked with red star for all CK+ sequences



Although the last frame is not always selected as the key frame, the selected key frames are very close to last frames. Besides this, the number of selected key frames, on average, is very close to one. At least one and at most three frames are selected as key frames (Figure 4.7 and Table 4.2).

Figure 4.7: Number of selected keyframes for CK+ sequences



The sequences are represented by one key frame on the average (Table 4.2). This ratio shows that actually a sequence of CK+ database may be represented by 5.8% of actual frames.

Table 4.2: Statistics of selected keyframe counts for CK+ sequences

Min	Max	Mean
1	3	1.05

We conducted facial expression recognition experiments using the selected key frames. Unfortunately, the available datasets that are used for affect recognition does not have many samples and conducting subject independent tests is an important issue in order to

ensure the reliability of the recognition accuracy. Therefore, while calculating the recognition accuracy, Leave-One-Subject-Out (LOSO) test strategy is used.

LOSO cross-validation involves using the data that belongs to one subject as the test set for each test cycle. The remaining data of other subjects are reserved for the training of the classification system as the training set. This test cycle is repeated until all subjects in the database are all tested. LOSO cross validation makes the experimental results subject independent since the subject being tested does not exist in the training set. The tests of CK+ database are conducted on the sequences that have emotion by preserving LOSO test strategy.

When all frames are taken into consideration and the average of LPQ features for all frame of one sequence is used as the feature of that sequence, the reached emotion recognition accuracy is 89.6% (Table 4.3). The emotion recognition accuracy results of our framework on CK+ database is 92.0% using Zhu's facial landmark detection method [147] and LPQ features of selected key frames. However, it is known that, for CK+ database, peak expressions are in the last frames of each sequence. So, intuitively, if the last frames are used as peak frames, the recognition accuracy becomes 91.4%.

Table 4.3: Facial recognition rates for the CK+ database

Used Frames	Facial Landmarks	Weighted Recognition Rate	Unweighted Recognition Rate
All frames	CK+	89.6%	84.6%
The last frame of each sequence	CK+	93.3%	89.1%
Selected key frames by our algorithm	CK+	95.1%	93.5%
The last frame of each sequence	Zhu	91.4%	83.9%
Selected key frames by our algorithm	Zhu	92.0%	88.1%

Some of the error comes from false detection of landmarks. Hence, if the landmarks which are given with CK+ databases are used instead of the landmarks tracked by Zhu's method [147] overall accuracy of the key frame selection based method increases to 95.1 percent. The mentioned recognition results are summarized in Table 4.3. The confusion matrices of the case using the landmarks given in CK+ database shown in Table 4.4 and Table 4.5, respectively.

Table 4.4: Confusion matrix for the CK+ database when the last frame (i.e. the peak frame) and CK+ landmarks are used

	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**93.3%**	0.0%	4.4%	0.0%	0.0%	2.2%	0.0%
Contempt	5.6%	**83.3%**	0.0%	0.0%	5.6%	0.0%	5.6%
Disgust	0.0%	0.0%	**98.3%**	0.0%	1.7%	0.0%	0.0%
Fear	0.0%	4.0%	0.0%	**76.0%**	12.0%	0.0%	8.0%
Happiness	0.0%	0.0%	0.0%	0.0%	**100.0%**	0.0%	0.0%
Sadness	17.9%	3.6%	0.0%	0.0%	0.0%	**75.0%**	3.6%
Surprise	1.2%	1.2%	0.0%	0.0%	0.0%	0.0%	**97.6%**

Table 4.5: Confusion matrix for the CK+ database using the selected key frames and CK+ landmarks

	Anger	Contempt	Disgust	ear	Happiness	Sadness	Surprise
Anger	**86.7%**	4.4%	4.4%	0.0%	0.0%	4.4%	0.0%
Contempt	0.0%	**94.4%**	0.0%	5.6%	0.0%	0.0%	0.0%
Disgust	0.0%	0.0%	**98.3%**	0.0%	1.7%	0.0%	0.0%
Fear	0.0%	0.0%	0.0%	**92.0%**	0.0%	0.0%	8.0%
Happiness	0.0%	0.0%	0.0%	0.0%	**100.0%**	0.0%	0.0%
Sadness	14.3%	0.0%	0.0%	0.0%	0.0%	**85.7%**	0.0%
Surprise	0.0%	1.2%	1.2%	0.0%	0.0%	0.0%	**97.6%**

We achieved a maximum facial expression recognition rate of 95.1 percent on the CK+ database. If it is compared with other methods in the literature (see Table 4.6), we can see that there are higher accuracies reported. We would like to note that in our frame work, we focused on the selection of the minimum number of key frames that represent the actual sequence as good as possible. We do not focus on the selection of the features that represent the sequence in the best way and give the highest recognition accuracy. We used LPQ features in our tests. It is possible that there might be other facial features that might give higher accuracies.

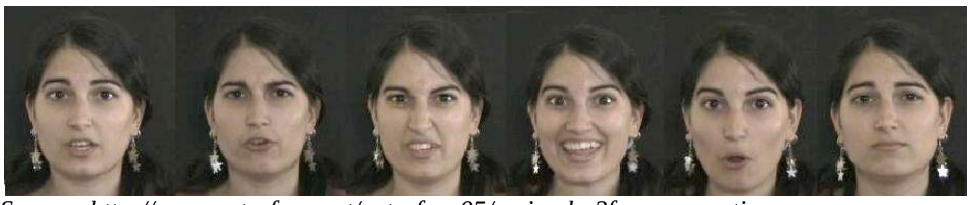
Table 4.6: Performances of current facial-expression recognition methods on CK+

Ref	Classes	Evaluation	Recognition Rate
Kotsia et. al [68]	7	5-fold	92.3%
Shan et. al [113]	7	10-fold	91.4%
Zisheng et. al [73]	6	LOSO	96.33%
Shojaeilangari et. al [114]	6	LOSO	92.97%
Gu et. al[53]	7	10-fold	91.51%
Xue et. al [91]	6	5-fold	89.2%
Ulukaya et al. [123]	7	LOSO	90%
Our framework	7	LOSO	95.1%

4.2.2 Results on eNTERFACE'05 Database

Another elicited audio-visual dataset is eNTERFACE'05 [88] that was collected during the eNTERFACE'05 workshop. It contains audio-visual clips of 42 subjects from 14 different nationalities. Among the subjects, a percentage of 81% were men, while the remaining 19% were women. A percentage of 31% of the total set wore glasses, while 17% of the subjects had a beard. The database was recorded using a standard mini-DV digital video camera that has 800.000 pixels resolution.

Figure 4.8: A sample sequence from eNTERFACE'05 database



Source: http://www.enterface.net/enterface05/main.php?frame=emotion

After each subject had listened to six different short stories that include basic emotional states as anger, disgust, fear, happiness, sadness and surprise, they uttered given sentences with the target emotion in English.

Table 4.7: Frame counts for eNTERFACE'05 sequences and selected key frames

	Min	Max	Mean
Frames in Actual Sequences	28	171	70
Selected Key frames	1	20	4.7

There are 1287 sequence in eNTERFACE'05 database which includes six basic emotions and they have 70 frames on the average. The minimum and maximum number of frames that a sequence may have and the distribution of the number of frames can be seen in Table 4.7 and Figure 4.9 respectively.

Figure 4.9: Distribution of the number of frames for eNTERFACE'05 sequences

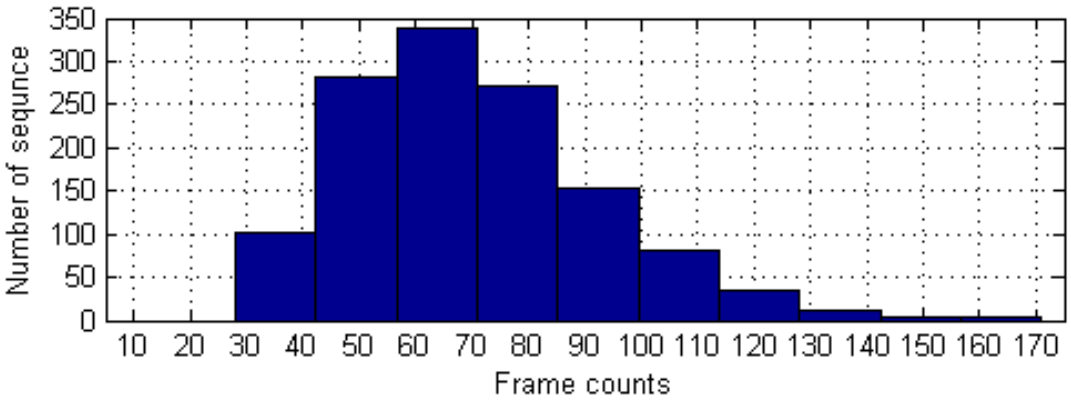
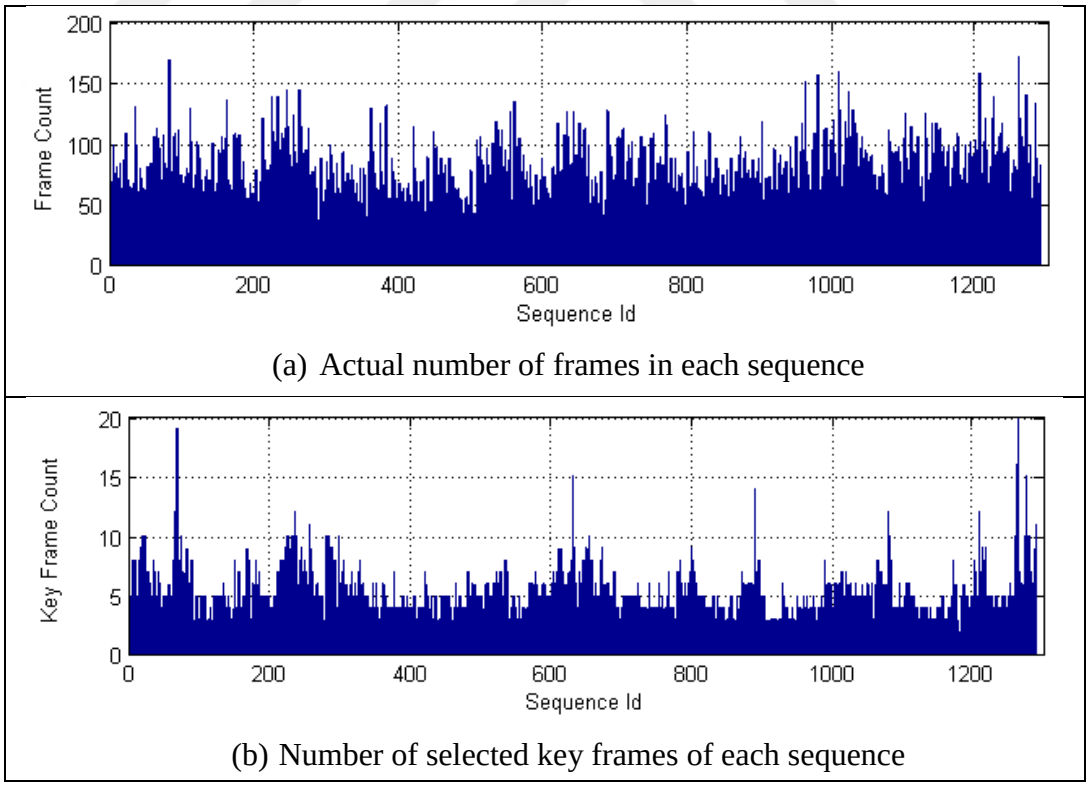


Figure 4.10: Number of frames in eNTERFACE'05 sequences



(a) Actual number of frames in each sequence

(b) Number of selected key frames of each sequence

When our key frame extraction algorithm is applied to the eNTERFACE'05 database, the sequences are represented by 20 key frames on the average (Table 4.7). This ratio shows that actually a sequence of eNTERFACE'05 may be represented by 28% of its actual frames. Distribution of actual and selected key frame counts for each sequence is drawn in Figure 4.10.

Similar to the CK+ database, in order to ensure subject independence on the eNTERFACE'05 database we used Leave-One-Subject-Out cross-validation strategy. When all frames are taken into consideration and the average of LPQ features for all frame of one sequence is used as the feature of that sequence, the reached LOSO test emotion recognition accuracy is 40.87%. By using only average of selected key frames' LPQ features, a video-based affect recognition accuracy of 43.82% is calculated (see When MFCC and RASTA-PLP features are used as audio features, 65.89% affect recognition accuracy is reached. In the case of some decision level fusion is applied to audio and selected key frame cases, calculated affect recognition accuracy is 62.70%.

Table 4.8 and Table 4.9). When MFCC and RASTA-PLP features are used as audio features, 65.89% affect recognition accuracy is reached. In the case of some decision level fusion is applied to audio and selected key frame cases, calculated affect recognition accuracy is 62.70%.

Table 4.8: Recognition rates for eNTERFACE'05 database

	Weighted Recognition Rate	**Unweighted Recognition Rate**
Single modalities		
LPQ of all frames	40.87%	43.33%
LPQ of selected key frames	43.82%	45.95%
Audio	65.89%	66.72%
Decision level fusion		
Sum Rule	61.07%	63.00%
Product Rule	62.70%	63.91%

If the result of the recognition accuracy is compared with other methods in the literature that are tested on eNTERFACE'05 database (see Table 4.10), it can be easily said that

key frame selection algorithm does not reduce, even increases the accuracy although it reduces the number of representative frame very much (one third for eNTERFACE'05).

Table 4.9: Confusion matrix for eNTERFACE'05 database when LPQ of selected key frames are used

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**32.6%**	11.2%	12.1%	12.1%	17.2%	14.9%
Disgust	7.4%	**63.7%**	5.6%	12.1%	6.5%	4.7%
Fear	20.0%	13.0%	**20.5%**	6.0%	23.3%	17.2%
Happiness	9.9%	7.1%	6.6%	**64.6%**	4.2%	7.5%
Sadness	13.0%	8.8%	13.0%	6.0%	**48.4%**	10.7%
Surprise	21.4%	3.3%	14.0%	14.4%	13.5%	**33.5%**

In our tests, we used basic visual features as LPQ and achieved higher than average performance in the literature. There are several methods that give higher accuracies than our method, but their cross-validation schemes do not guarantee subject independent results. It shows the power of key frame selection based on minimum sparse reconstruction.

Table 4.10: Performances of expression recognition methods on eNTERFACE'05

Ref	Number Of Subjects	Evaluation	Recognition Rate (%)
M. Paleari [99]	44	No Info.	25.0
M. Mansoorizadeh [86]	42	10-fold	37.0
R. Gajsek [49]	No Info.	5-fold	54.7
D. Datcu [31]	42	3-fold	37.7
Y. Wang et. Al [133]	43	10-fold	58
H. Kuan-Chieh et. Al [60]	No Info.	6-fold	52.3
Our approach	43	LOSO	43.8

In 2015, Zhalehpour [139] [140][141] presented several peak frame detection methods. She also used eNTERFACE'05 database and analyzed the performance of different key frame selection methods on the affect recognition accuracy. These peak frame selection methods are maximum dissimilarity, emotion intensity, and clustering based methods. She compared these techniques with manually selected peak frames' recognition accuracy. She achieved 40.00% affect recognition accuracy with clustering based automatic peak frame selection technique against 47.05 percent with manually selected

ones. Our minimum sparse reconstruction based automatic key frame selection algorithm achieved a higher recognition rate which is 43.82 percent (see Table 4.11).

Table 4.11: Peak frame selection methods on eNTERFACE'05 database.

Peak/Key Frame Selection Method	Recognition Rate (%)
Manual Frame Selection [139]	47.05
Maximum Dissimilarity based [139]	38.22
Emotion Intensity based [139]	39.38
Clustering based [139]	40.00
Audio based [139]	34.46
Minimum Sparse Reconstruction based method	43.82

4.2.3 Results on BAUM-1a Database

BAUM-1 (Bahçeşehir University Multimodal Affective Database - 1) [98] [44] is a collection of audio-visual facial clips of acted and spontaneous (re-acted) affective expressions. The audio-visual clips have been recorded from 31 subjects, who express a rich set of emotional and mental states in an unscripted way in Turkish. The database contains synchronous facial recordings of subjects with a frontal stereo camera and a half profile mono camera.

The subjects first watch visual or audio-visual stimuli on a screen in front of them, which are designed and timed to elicit certain emotions and mental states. The subjects answer questions and express their feelings about the visual stimuli in their own words. The target emotions that have been elicited are the five basic ones (happiness, anger, sadness, disgust, fear) and additionally boredom and contempt. There are also several mental states including unsure (such as confusion, undecidedness), thinking, concentration, interest (including curiosity), and bothered (inc. complaint). The database also contains short acted recordings of each subject. The video clips have been categorically annotated by five labelers. Also a score between 0-5 is given to each video clip indicating the activation level at the peak frame of the emotion or mental state expressed in the video clip.

BAUM-1a database is collected in an acted way where the subjects are uttering some predefined sentences with the target emotions. Due to differences between the length of the sentences and the time duration of the speech for every subject, the length of the video clips is changing between one and sixteen seconds.

Table 4.12: Properties of the BAUM-1 database

Feature	Acted	Spontaneous
Number of Video Clips	280	1222
Number of Subjects	31	
Male / Female Ratio	18/13	
Age Range	18 - 66	
KAPPA Value	0.64	0.54
Number of Videos per Emotion / Mental State	Happiness: 27 Sadness: 39 Anger: 43 Disgust: 35 Fear: 36 Surprise: 2 Boredom: 30 Interest: 30 Unsure: 38	Happiness: 161 Sadness: 148 Anger: 94 Disgust: 110 Fear: 52 Surprise: 54 Boredom: 43 Contempt: 19 Interest: 21 Unsure: 128 Neutral: 159 Bothered: 74 Concentrating: 61 Thinking: 98

Source: http://baum1.bahcesehir.edu.tr/

The video clips also include lots of lip and mouth movements, some head pose changes, as well as translational and rotational head motions. As a result of all these frame varieties in the sequences, the extraction of a unique or at least few representative key frames becomes a tough problem.

In order to investigate the database in details, the tests are conducted on two separate emotion categories as 5 basic emotions case (Anger, Disgust, Fear, Happiness and Sadness) and 8 emotions case (Boredom, Interest and Unsure are added to 5 basic emotions). The same test strategy (Leave-One_Subject-Out) is used in BAUM-1a 5 and 8 emotions database tests.

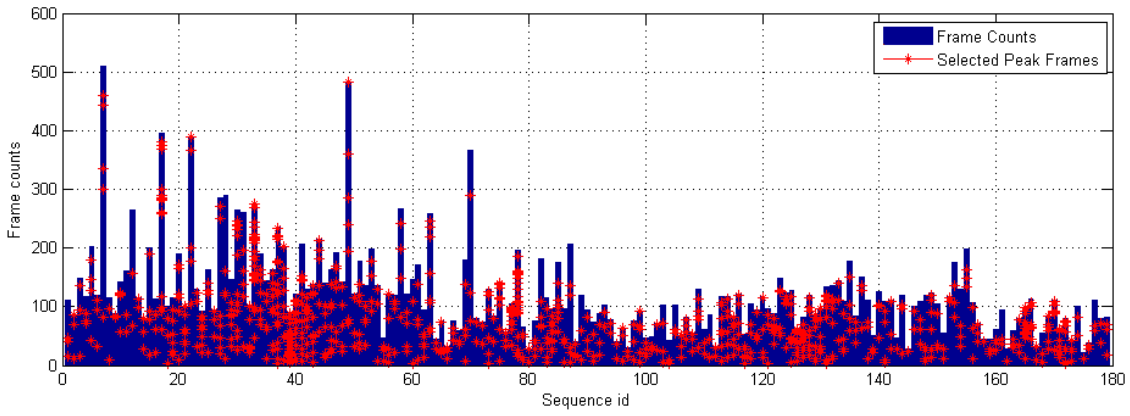
In BAUM-1a 5 emotion case, there are 179 sequences. These sequences have a minimum and maximum number of frames as 21 and 509, respectively (Table 4.13). The average number of frames in the sequences is 123.

Table 4.13: Actual and extracted number of key frames for BAUM-1a 5 emotion case

	Min	Max	Mean
Frames in Actual Sequences	21	509	123
Selected Key frames	1	25	4.8

When our key frame extraction algorithm is applied to BAUM-1a database, the sequences are represented by 4.8 key frames on the average (see Table 4.13). This ratio shows that actually a sequence of BAUM-1a database may be represented by 4% of its actual frames. In contradiction to the CK+ database, the intensity of the emotions in BAUM-1a database does not start as neutral in the first frame and increase while frames proceed. The intensity of the emotion may reach its peak anywhere during the clip. If the distribution of the location of the selected key frames is investigated in Figure 4.11, it is seen that the selected key frames are located homogeneously for each sequence. The selections do not concentrate on any specific location.

Figure 4.11: Key frame locations are marked for BAUM-1a database for 5 emotions



Similar to the CK+ and eNTERFACE'05 databases, LOSO cross validation tests are conducted. When all frames are taken into consideration and the average of LPQ features for all frames of one sequence is used as the feature of that sequence, the reached LOSO test emotion recognition accuracy is 61.45%. The test results showed an affect recognition accuracy of 60.89% by using our key frame selection algorithm (see Table 4.14 and Table 4.15). When MFCC and RASTA-PLP features are used as audio features, 78.21% affect recognition accuracy is reached. In the case of decision level

sum rule is applied to audio and selected key frame cases, calculated affect recognition accuracy is 81.56%.

Table 4.14: Recognition rates for the BAUM-1a database for 5 emotions

	Weighted Recognition Rate	Unweighted Recognition Rate
Single modalities		
LPQ of all frames	61.45%	62.11%
LPQ of selected key frames	60.89%	62.39%
Audio	78.21%	77.63%
Decision level fusion		
Sum Rule	81.56%	81.56%
Product Rule	81.01%	80.50%

Table 4.15: Confusion matrix for 5 basic emotions using BAUM-1a database when LPQ of selected key frames are used.

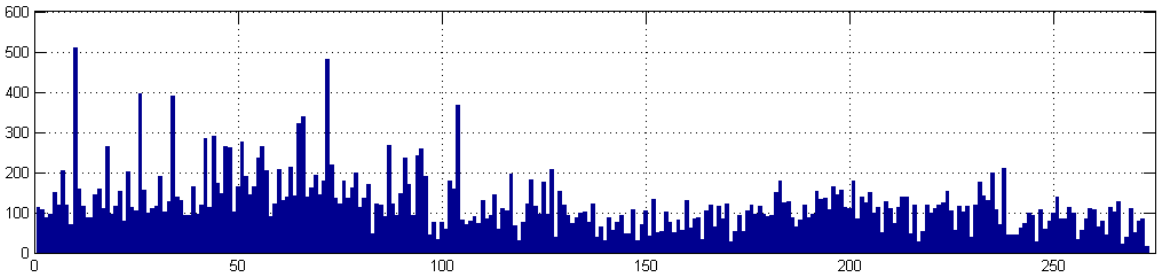
	Anger	Disgust	Fear	Happiness	Sadness
Anger	**48.8%**	9.3%	14.0%	9.3%	18.6%
Disgust	5.7%	**77.1%**	0.0%	11.4%	5.7%
Fear	22.2%	0.0%	**55.6%**	2.8%	19.4%
Happiness	7.4%	3.7%	11.1%	**77.8%**	0.0%
Sadness	13.2%	15.8%	15.8%	2.6%	**52.6%**

In BAUM-1a database for the 8 emotion case, there are 273 sequences. The sequences that have a minimum and maximum number of frames are 17 and 509 frames, respectively (see Table 4.16). The average number of frames in the sequences is 122. The distribution of the number of frames for each sequence in BAUM-1a database can be seen in Figure 4.12

Table 4.16: Frame counts for BAUM-1a 8 emotion case

	Min	Max	Mean
Frames in Actual Sequences	17	509	122
Selected Key frames	1	25	4.8

Figure 4.12: Frame counts of BAUM-1a 8 emotion



When LOSO tests are conducted on BAUM-1a database, in the case of 8 emotions are included, the test results showed 41.76% affect recognition accuracy when all frames are taken into consideration and the average of LPQ features for all frames of one sequence is used as the feature of that sequence. By using our key frame selection algorithm, we reached 36.63% affect recognition accuracy (see Table 4.17 and Table 4.18).

Table 4.17: Recognition rates for the BAUM-1a 8 emotion

	Weighted Recognition Rate	Unweighted Recognition Rate
Single modalities		
LPQ of all frames	41.76%	40.62%
LPQ of selected key frames	36.63%	35.96%
Audio	65.93%	65.96%
Decision level fusion		
Sum Rule	67.77%	67.74%
Product Rule	68.50%	67.89%

Table 4.18: Confusion matrix for the 8 basic emotions using BAUM-1a when LPQ of selected key frames are used.

	Anger	Boredom	Disgust	Fear	Happiness	Interest	Sadness	Unsure
Anger	**41.9%**	0.0%	14.0%	9.3%	14.0%	0.0%	18.6%	2.3%
Boredom	29.6%	**0.0%**	3.7%	29.6%	0.0%	3.7%	18.5%	14.8%
Disgust	5.7%	0.0%	**80.0%**	0.0%	11.4%	0.0%	2.9%	0.0%
Fear	27.8%	0.0%	0.0%	**27.8%**	2.8%	5.6%	16.7%	19.4%
Happiness	11.1%	0.0%	7.4%	7.4%	**74.1%**	0.0%	0.0%	0.0%
Interest	24.1%	0.0%	6.9%	20.7%	20.7%	**3.4%**	13.8%	10.3%
Sadness	18.4%	0.0%	15.8%	10.5%	0.0%	2.6%	**50.0%**	2.6%
Unsure	39.5%	0.0%	5.3%	18.4%	5.3%	2.6%	18.4%	**10.5%**

When MFCC and RASTA-PLP features are used as audio features, 65.93% affect recognition accuracy is reached. In the case of decision level product rule is applied to audio and selected key frame cases, calculated affect recognition accuracy is 68.50% (see Table 4.17).

Similar to eNTERFACE'05 database, Zhalehpour [139] also performed same tests on BAUM-1a database. For 5 emotion case, she achieved 55.70% affect recognition accuracy with clustering based automatic peak frame selection technique against 55.61% with manually selected ones. Our minimum sparse reconstruction based automatic key frame selection algorithm achieved higher recognition rate with 60.89% for 5 emotion case and 36.63% for 8 emotion case (see Table 4.21), which shows an improvement of almost 5% over the results in [139].

Table 4.19: Peak frame selection methods on BAUM-1a.

Peak Frame Selection Method	5 Emotion Recognition Rate (%)	8 Emotion Recognition Rate (%)
Manual Frame Selection [139]	55.61	31.32
Maximum Dissimilarity based [139]	46.60	26.30
Emotion Intensity based [139]	52.06	29.55
Clustering based [139]	55.70	36.33
Audio based [139]	42.18	20.83
Minimum Sparse Reconstruction based	60.89	36.63

4.2.4 Emotion Recognition in the Wild (EmotiW 2015) Challenge

We also participated in the third Emotion Recognition in the Wild Challenge (EmotiW 2015) at ACM Int. Conf. on Multimedia Interaction (ICMI) which mimics real-world conditions. The EmotiW 2015 challenge is based on the AFEW 5.0 (Acted Facial Expressions in the Wild) database, which contains short audio-visual clips collected from movies and labeled using a semi-automatic approach described in [34]. AFEW 5.0 is divided into three parts for training, validation and testing. The numbers of samples for each emotion in each set are shown in Table 4.20. The labels of the test set are not given to the participants of the challenge.

Table 4.20: The numbers of samples in EmotiW 2015 AFEW database.

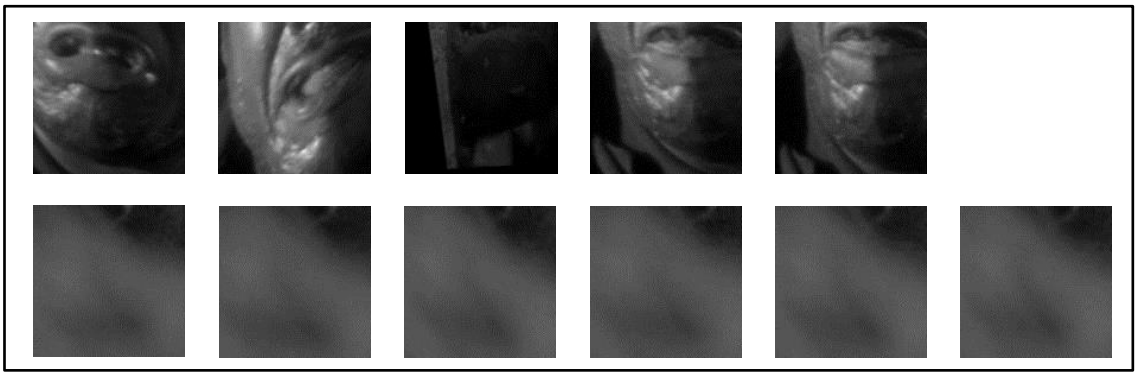
	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Train	118	72	77	145	131	107	73
Validation	64	40	46	63	63	61	46
Test	79	29	66	108	159	71	27

The organizers of the EmotiW 2015 challenge provide the LBP-TOP feature set on the AFEW 5.0 database. After the pre-processing step for aligning and cropping the face region, LBP-TOP features are extracted from non-overlapping spatial 4x4 blocks. The LBP-TOP features from each block are concatenated to create one feature vector of size 1×2832 for each frame of a video.

The video only baseline classification accuracy provided by the challenge organizers using an SVM classifier with Chi square kernel on the validation set is 36.08%. Similarly, baseline classification accuracy on the test set is 39.33%.

We would like to note that while the database comes with Zhu's [147] face tracking results, the difficult and close-to-real-world conditions cause problems even in the early stages of the processing pipeline. Some examples of incorrectly detected and tracked face images are illustrated in Figure 4.13.

Figure 4.13: Noisy face detection and tracking results from AFEW 5.0 dataset.



We evaluated the performance of video features (LBP-TOP and LPQ`s from key frames) and audio features (OpenSMILE and MFCC & RASTA-PLP features) on the validation and test sets. The emotion recognition results of eight combinations on validation and test sets are illustrated in Table 4.21. The configuration used for each test case is explained below.

a. **Case 1:** Video based experiment, where key frames of all video sequences are selected using T_P as 0.80. For each video sequence, average LPQ feature vector of all key frames is used to represent the whole video. An SVM with a chi-square kernel is trained.
b. **Case 2:** Video based experiment combining case 1 and a second chi-square kernel SVM, which is trained using LBP-TOP features. Then score level fusion is applied using the product rule.
c. **Case 3:** Video based experiment, which is same as case 2 but the parameter T_P has been selected as 0.85.
d. **Case 4:** Same as case 3 with the SVM classifier trained using both the training and the validation sets of AFEW 5.0 dataset.
e. **Case 5:** Audio based experiment using feature level fusion of MFCC and RASTA-PLP features. An SVM classifier with an exponential chi-square kernel is used.
f. **Case 6:** Audio based experiment using the OpenSMILE feature set. An SVM classifier with an exponential chi-square kernel is used.
g. **Case 7:** Audio-visual experiment using score level fusion of Case 2 and Case 5.
h. **Case 8:** Audio-visual experiment using score level fusion of Case 2 and Case 6.

Table 4.21: Accuracies of the 8 test cases on validation and test sets.

Case	Methods		Accuracy	
			Val. (%)	Test (%)
Video Based Baseline			36.08	39.33
1	Video Based	LPQ ($T_P = 0.80$)	40.70	41.37
2		LBP-TOP + LPQ ($T_P = 0.80$)	43.40	45.08
3		LBP-TOP + LPQ ($T_P = 0.85$)	44.47	44.71
4		LBP-TOP + LPQ (trained by train+val set) ($T_P = 0.80$)	-	43.23
5	Audio Based	MFCC & RASTA-PLP	24.02	33.40
6		OpenSMILE	31.85	33.21
7	Video + Audio Based	LBP-TOP + LPQ + MFCC & RASTA-PLP	41.24	47.68
8		LBP-TOP + LPQ + OpenSMILE	40.70	49.91

The best classification accuracy on the test set is achieved for Case 8, using score level fusion of LBP-TOP, key frame LPQ's and OpenSMILE audio features, which is 49.91%. The confusion matrices of this case for the validation and test sets are shown in Table 4.22 and Table 4.23, respectively.

Table 4.22: Confusion matrix of Case 8 (audio-visual) on validation set.

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	**67.8%**	0.00%	0.00%	13.5%	6.78%	8.47%	3.39%
Disgust	20.5%	**7.69%**	0.00%	23.08%	25.64%	17.95%	5.13%
Fear	34.0%	0.00%	**11.3%**	15.91%	20.45%	9.09%	9.09%
Happiness	4.7%	1.59%	0.00%	**79.37%**	9.52%	4.76%	0.00%
Neutral	4.92%	0.00%	4.92%	26.23%	**55.74%**	8.20%	0.00%
Sadness	11.8%	3.39%	8.47%	18.64%	32.20%	**25.42%**	0.00%
Surprise	28.2%	2.17%	10.8%	13.04%	30.43%	6.52%	**8.70%**

Table 4.23: Confusion matrix of Case 8 (audio-visual) on test set.

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	**73.42%**	1.27%	3.80%	6.33%	12.66%	2.53%	0.00%
Disgust	13.79%	**0.00%**	0.00%	37.93%	17.24%	20.69%	10.34%
Fear	31.82%	3.03%	**18.18%**	7.58%	10.61%	16.67%	12.12%
Happiness	5.56%	0.93%	0.93%	**73.15%**	8.33%	9.26%	1.85%
Neutral	5.03%	2.52%	0.63%	15.72%	**55.35%**	18.87%	1.89%
Sadness	12.68%	4.23%	2.82%	21.13%	14.08%	**42.25%**	2.82%
Surprise	22.22%	0.00%	7.41%	14.81%	29.63%	18.52%	**7.41%**

We can see that happiness and anger have the highest accuracies. In Table 4.24 and Table 4.25, we give the confusion matrices of Case 2 (video only) and Case 6 (audio only), as well.

Table 4.24: Confusion matrix of Case 2 (Visual) on test set.

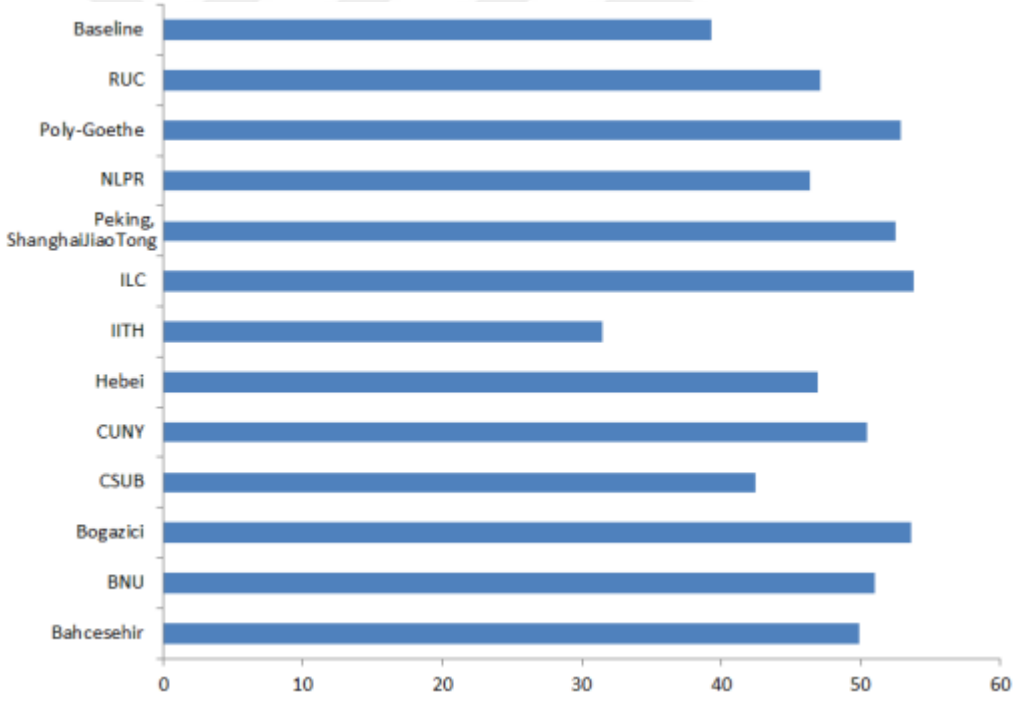
	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	**73.42%**	2.53%	5.06%	6.33%	6.33%	6.33%	0.00%
Disgust	10.34%	**17.24%**	3.45%	17.24%	13.79%	24.14%	13.79%
Fear	27.27%	3.03%	**15.15%**	6.06%	15.15%	18.18%	15.15%
Happiness	9.26%	2.78%	0.93%	**68.52%**	4.63%	12.04%	1.85%
Neutral	10.69%	3.14%	3.14%	15.72%	**43.40%**	18.87%	5.03%
Sadness	12.68%	7.04%	5.63%	22.54%	11.27%	**33.80%**	7.04%
Surprise	37.04%	0.00%	7.41%	14.81%	18.52%	11.11%	**11.11%**

Table 4.25: Confusion matrix of Case 6 (Audio) on test set.

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	**55.70%**	0.00%	11.39%	20.25%	7.59%	3.80%	1.27%
Disgust	13.79%	**0.00%**	3.45%	31.03%	34.48%	17.24%	0.00%
Fear	25.76%	1.52%	**18.18%**	21.21%	21.21%	10.61%	1.52%
Happiness	14.81%	0.93%	0.00%	**43.52%**	25.93%	12.04%	2.78%
Neutral	6.29%	0.63%	5.66%	33.33%	**36.48%**	16.35%	1.26%
Sadness	15.49%	0.00%	5.63%	25.35%	26.76%	**23.94%**	2.82%
Surprise	11.11%	3.70%	3.70%	18.52%	33.33%	25.93%	**3.70%**

In total, 75 teams registered for the challenge and 22 papers were submitted. Figure 4.14 shows the classification accuracy performance of challenge participants on test set. We achieved 49.91% classification accuracy on test set while base line accuracy was 39.33%. Our audio-visual affect recognition accuracy was the 7th among 13 teams which submitted their results.

Figure 4.14: EmotiW 2015 challenge results on test sets



Source:[35]

5. CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

In this thesis, we investigated audio-visual affect recognition from video clips. Our approach was based on choosing the key frames that represent the content of the whole video clip in the best possible way so that the emotional content of this video can be recognized more efficiently. We have adopted a new video summarization method [89] to our problem and proposed it as minimum sparse reconstruction based key frame selection in video for affect recognition. We extracted LPQ and LBP-TOP video features and selected key frames by investigating the sparsest representative subset of them. Besides visual data, we also extracted MFCC and RASTA-PLP audio features and fused them with video at the score level using the multiplication rule.

We tested the proposed key frame selection algorithm on three widely used visual databases (CK+, eNTERFACE and BAUM1-a). Zhalehpour [139] conducted similar research on several peak frame selection methods using eNTERFACE and BAUM1-a databases. The work in [139] focused on maximum dissimilarity, emotion intensity, clustering and audio based methods. The highest recognition accuracies by using a clustering based key frame extraction method have been reported as 40.00 and 36.33 percent on the eNETRAFACE and BAUM-1a databases in [139],[141], respectively. Our proposed method achieved respectively 44.72 and 39.13 percent recognition accuracies, which are higher than the results reported in [141] and are very close to the accuracies that have been reported by manually selecting the key frames in [141], which can be considered as ground-truth. Furthermore, we attended a worldwide challenge (EmotiW 2015) to test our proposed method on a more challenging and realistic database. The database of the challenge consists of audio-visual clips collected from movies and imitates real world conditions. The affect recognition accuracy of challenge baseline is 39.33 percent and we reached to 49.91 percent. Our method was the 7th among 13 teams which submitted their results.

The results of our proposed method are promising. Conducted tests showed that by efficiently selecting key frames of a video, it is possible to represent the affective content of video without compromising recognition accuracy.

5.2 FURTHER RESEARCH

In this thesis, we mainly focused on the selection of key frames for an affective facial video content. We proposed a sparse reconstruction based method that looks for the most sparse and most representative subset of the frames among whole sequences. We used state-of-the-art LPQ and LBP-TOP visual features and did not investigate the best features that can be used. It is obvious that using some other features and learning techniques affect recognition accuracy may be increased. As can be seen from EmotiW 2015 challenge results, a large number of the proposed methods used deep learning based techniques and by considering the recognition accuracies it can be said that deep learning based methods are promising.

Additionally, the performance of the methods is limited by the amount of labeled data and performance limitations of the current state-of-art face and facial parts detectors. Therefore there is need for such databases that contains more labeled and realistic data. Also face and facial parts detection is another independent research area that will affect the recognition accuracy directly.

Furthermore, by considering the database that we used in the challenge, which is closer to the real world conditions, some scenes, frames or even audio may contain multiple subjects who express multiple emotions or multiple emotion intensities. This in itself is a challenging problem and need to be further looked into.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 12. pp. 2037–2041, 2006.

[2] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using Local Phase Quantization," *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. pp. 1–4, 2008.

[3] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face Recognition with Local Binary Patterns," in *Computer Vision - ECCV 2004 SE - 36*, vol. 3021, T. Pajdla and J. Matas, Eds. Springer Berlin Heidelberg, 2004, pp. 469–481.

[4] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychol. Bull.*, vol. 111, no. 2, pp. 256–274, 1992.

[5] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proceedings of ICSLP*, 2002, pp. 2037–2040.

[6] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimed. Syst.*, vol. 16, no. 6, pp. 345–379, Nov. 2010.

[7] A. Bafandehkar, M. Nazari, and M. Rahat, "Pictorial structure based keyparts localization for facial expression recognition using Gabor filters and Local Binary Patterns Operator," in *2011 International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 2011, pp. 429–434.

[8] M. R. Banham and a K. Katsaggelos, "Digital image restoration," *Signal Process. Mag. IEEE*, vol. 14, no. 2, pp. 24–41, 1997.

[9] A. Batliner, C. Hacker, S. Steidl, and E. Nöth, "'You Stupid Tin Box'-Children Interacting with the AIBO Robot: A Cross-linguistic Emotional Speech Corpus.," *Lrec*, pp. 171–174, 2004.

[10] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus.," *Proc. Work. Corpora Res. Emot. Affect Lr.*, pp. 28–31, 2008.

[11] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Commun.*, vol. 40, no. 1–2, pp. 117–143, Apr. 2003.

[12] B. B. Bauer, "The Measurement of Loudness Level," *J. Acoust. Soc. Am.*, vol. 50, no. 2A, p. 405, 1971.

[13] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem, "Use of Line Spectral Frequencies for Emotion Recognition from Speech," *2010 20th Int. Conf. Pattern Recognit.*, pp. 3708–3711, 2010.

[14] E. Bozkurt, E. Erzin, Ç. E. Erdem, and A. T. Erdem, "Formant position based weighted spectral features for emotion recognition," *Speech Commun.*, vol. 53, no. 9–10, pp. 1186–1197, 2011.

[15] F. Burkhardt, a Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *Ninth Eur. Conf. Speech Commun. Technol.*, vol.

2005, pp. 3–6, 2005.

[16] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, “Analysis of Emotion Recognition using Facial Expressions , Speech and Multimodal Information,” pp. 1–7, 2004.

[17] P. Case, “Emotion Recognition Based on MFCC Features using SVM,” vol. 7782, pp. 31–36, 2014.

[18] Ce Liu, J. Yuen, and A. Torralba, “SIFT Flow: Dense Correspondence across Scenes and Its Applications,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[19] C.-C. Chang and C.-J. Lin, “Libsvm,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[20] K. Y. Chang, T. L. Liu, and S. H. Lai, “Learning partially-observed hidden conditional random fields for facial expression recognition,” *2009 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work. 2009*, pp. 533–540, 2009.

[21] Y. Chavhan, M. L. Dhore, and P. Yesaware, “Speech Emotion Recognition using Support Vector Machine,” *Int. J. Comput. Appl.*, vol. 1, no. 20, pp. 8–11, 2010.

[22] C.-Y. Chen, Y.-K. Huang, and P. Cook, “Visual/Acoustic Emotion Recognition,” *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. pp. 1468–1471, 2005.

[23] L. S. Chen, H. Tao, T. S. Huang, T. Miyasato, and R. Nakatsu, “Emotion recognition from audiovisual information,” in *1998 IEEE Second Workshop on Multimedia Signal Processing (Cat. No.98EX175)*, 1998, no. 1, pp. 83–88.

[24] X. Cheng and Q. Duan, “Speech Emotion Recognition Using Gaussian Mixture Model,” *Int. Conf. Comput. Appl. Syst. Model.*, pp. 1222–1225, 2012.

[25] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, “Facial expression recognition from video sequences: Temporal and static modeling,” *Comput. Vis. Image Underst.*, vol. 91, no. 1–2, pp. 160–187, 2003.

[26] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[27] S. F. Cotter, “Sparse Representation for accurate classification of corrupted and occluded facial expressions,” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 838–841.

[28] R. Cowie, J. G. Taylor, S. Ioannou, M. Wallace, and S. Kollias, “An Intelligent System for Facial Emotion Recognition,” *Comput. Eng.*, pp. 0–3, 2005.

[29] A. Cruz, B. Bhanu, and S. Yang, “A psychologically-inspired match-score fusion model for video-based facial expression recognition,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6975 LNCS, no. PART 2, pp. 341–350, 2011.

[30] C. Darwin, *The expression of the emotions in man and animals*. United Kingdom: John Murray, 1872.

[31] D. Datcu and L. J. M. Rothkrantz, “Emotion recognition using bimodal data

fusion," in *Proceedings of the 12th International Conference on Computer Systems and Technologies - CompSysTech '11*, 2011, p. 122.

[32] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoust. Speech Signal Process. IEEE Trans.*, vol. 28, no. 4, pp. 357–366, 1980.

[33] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," *2011 IEEE Int. Conf. Autom. Face Gesture Recognit. Work. FG 2011*, pp. 878–883, 2011.

[34] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimed.*, vol. 19, no. 3, pp. 34–41, 2012.

[35] A. Dhall, O. V. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: EmotiW 2015," in *ACM International Conference on Multimodal Interaction (ICMI 2015)*, 2015.

[36] Dong-chen He and Li Wang, "Texture Unit, Texture Spectrum, And Texture Analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 4, pp. 509–512, Jul. 1990.

[37] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Commun.*, vol. 40, no. 1–2, pp. 33–60, 2003.

[38] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A New Emotion Database: Considerations, Sources and Scope," *In*, pp. 39–44, 2000.

[39] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data," *Affect. Comput. Intell. Interact.*, vol. 4738, pp. 488–500, 2007.

[40] Z. Duric, W. D. Gray, R. Heishman, Fayin Li, A. Rosenfeld, M. J. Schoelles, C. Schunn, and H. Wechsler, "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," *Proc. IEEE*, vol. 90, no. 7, pp. 1272–1289, Jul. 2002.

[41] P. Ekman and W. V Friesen, "Facial action coding system: a technique for the measurement of facial movement," 1978.

[42] P. Ekman, "Facial Expressions," *Handb. Cogn. Emot.*, vol. 16, pp. 301–320, 1999.

[43] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: The Manual on CD ROM*. Salt Lake City: A Human Face, 2002.

[44] C. E. Erdem, "http://baum1.bahcesehir.edu.tr," 2015. [Online]. Available: http://baum1.bahcesehir.edu.tr.

[45] C. E. Erdem, E. Bozkurt, E. Erzin, and a. T. Erdem, "RANSAC-based training data selection for emotion recognition from spontaneous speech," *Proc. 3rd Int. Work. Affect. Interact. Nat. Environ. - Affin. '10*, p. 9, 2010.

[46] C. Eroglu Erdem, C. Turan, and Z. Aydin, "BAUM-2: a multilingual audio-visual

affective face database," *Multimed. Tools Appl.*, pp. 7429–7459, 2014.

[47] H. Freeman, "On the Encoding of Arbitrary Geometric Configurations," *Electronic Computers, IRE Transactions on*, vol. EC-10, no. 2. pp. 260–268, 1961.

[48] D. Gabor, "Theory of communication. Part 3: Frequency compression and expansion," *J. Inst. Electr. Eng. - Part III Radio Commun. Eng.*, vol. 93, no. 26, pp. 445–457, Nov. 1946.

[49] R. Gajšek, V. Štruc, and F. Mihelič, "Multi-modal Emotion Recognition Using Canonical Correlations and Acoustic Features," *Pattern Recognit. (ICPR), 2010 20th Int. Conf.*, no. i, pp. 4141–4144, 2010.

[50] D. Galati, K. R. Scherer, and P. E. Ricci-Bitti, "Voluntary facial expression of emotion: comparing congenitally blind with normally sighted encoders.," *J. Pers. Soc. Psychol.*, vol. 73, no. 6, pp. 1363–1379, Dec. 1997.

[51] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining Prosodic Lexical and Cepstral Systems for Deceptive Speech Detection," in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, 2006, vol. 1, pp. I–1033–I–1036.

[52] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning, "Local features based facial expression recognition with face registration errors," in *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, 2008, pp. 1–8.

[53] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognit.*, vol. 45, no. 1, pp. 80–91, Jan. 2012.

[54] H. Gunes and M. Piccardi, "A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior," in *18th International Conference on Pattern Recognition (ICPR'06)*, 2006, vol. 1, pp. 1148–1153.

[55] H. Hermansky and N. Morgan, "RASTA processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4. pp. 578–589, 1994.

[56] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[57] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, and A. Stolcke, "Distinguishing Deceptive from Non-Deceptive Speech," in *Proceedings of Interspeech 2005*, 2005, pp. 1833–1836.

[58] C.-T. Hsu, S.-C. Hsu, and C.-L. Huang, "Facial Expression Recognition Using Hough Forest," *2013 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, pp. 1–9, 2013.

[59] T. Hu, L. C. De Silva, and K. Sengupta, "A hybrid approach of NN and HMM for facial emotion classification," *Pattern Recognit. Lett.*, vol. 23, no. 11, pp. 1303–1310, 2002.

[60] K. C. Huang, H. Y. S. Lin, J. C. Chan, and Y. H. Kuo, "Learning collaborative

decision-making parameters for multimodal emotion recognition," *Proc. - IEEE Int. Conf. Multimed. Expo*, 2013.

[61] K.-C. Huang, S.-Y. Huang, Y.-H. Kuo, Ò. Àù, Ù. À. Û. Ò. Ã. Ù. Ó. Ò. Åóò, and Ó. Ò. Àóòò, "Emotion recognition based on a novel triangular facial feature extraction method," *Neural Networks (IJCNN), 2010 Int. Jt. Conf.*, pp. 1–6, 2010.

[62] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. De la Torre, "Continuous AU intensity estimation using localized, sparse facial feature space," *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Work.*, pp. 1–7, 2013.

[63] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modeling," *IEEE Trans. Cybern.*, vol. 44, no. 2, pp. 161–174, Feb. 2014.

[64] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: different channels, same code?," *Psychol. Bull.*, vol. 129, no. 5, pp. 770–814, 2003.

[65] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, R. Memisevic, P. Vincent, A. Courville, and Y. Bengio, "Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video," *Proc. 15th ACM Int. Conf. Multimodal Interact.*, pp. 543–550, 2013.

[66] T. Kanade, J. F. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *Proc. Fourth IEEE Int. Conf. Autom. Face Gesture Recognit. (Cat. No. PR00580)*, pp. 46–53, 2000.

[67] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *Int. J. Hum. Comput. Stud.*, vol. 65, no. 8, pp. 724–736, Aug. 2007.

[68] I. Kotsia, S. Zafeiriou, N. Nikolaidis, and I. Pitas, "Texture and shape information fusion for facial action unit recognition," *Proc. 1st Int. Conf. Adv. Comput. Interact. ACHI 2008*, pp. 77–82, 2008.

[69] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion Recognition by Speech Signals," in *Eighth European Conference on Speech Communication and Technology*, 2003, pp. 125–128.

[70] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01.*, 2001, pp. 240–243.

[71] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, 2005.

[72] Li Fei-Fei and Pietro Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 524–531, 2005.

[73] Z. Li, J. Imai, and M. Kaneko, "Face and expression recognition based on bag of words method considering holistic and local image features," in *2010 10th International Symposium on Communications and Information Technologies*, 2010, pp. 1–6.

[74] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M. J. Rosato, "A 3D Facial Expression Database For Facial Behavior Research," in *7th International*

Conference on Automatic Face and Gesture Recognition (FGR06), 2006, vol. 2006, pp. 211–216.

[75] J. Liscombe, J. Hirschberg, and J. J. Venditti, "Detecting certainness in spoken tutorial dialogues," in *Interspeech*, 2005, pp. 1837–1840.

[76] C. L. Lisetti and F. Nasoz, "MAUI: a Multimodal Affective User Interface," in *Proceedings of the tenth ACM international conference on Multimedia - MULTIMEDIA '02*, 2002, p. 161.

[77] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1797–1803, 2009.

[78] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild," *Proc. 16th Int. Conf. Multimodal Interact. - ICMI '14*, pp. 494–501, 2014.

[79] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kande dataset (CK+): A complete facial expression dataset for action unit and emotionspecified expression," *IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, no. July, pp. 94–101, 2010.

[80] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews, "Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database," *Image Vis. Comput.*, vol. 30, no. 3, pp. 197–205, Mar. 2012.

[81] S. Lucey, A. B. Ashraf, and J. F. Cohn, "Investigating spontaneous facial action recognition through AAM representations of the face," 2005.

[82] D. Lundqvist, A. Flykt, and A. Ohman, "The Karolinska Directed Emotional Faces - KDEF," *CD ROM from Department of Clinical Neuroscience*. Psychology section, Karolinska Institutet, pp. 1–2, 1998.

[83] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," *Proc. - 3rd IEEE Int. Conf. Autom. Face Gesture Recognition, FG 1998*, pp. 200–205, 1998.

[84] L. Maat and M. Pantic, "Gaze-X: Adaptive, Affective, Multimodal Interface for Single-User Office Scenarios," in *Artifical Intelligence for Human Computing*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 251–271.

[85] K. Maekawa, "Phonetic and phonological characteristics of paralinguistic information in spoken japanese," *Icslp*, pp. 1–4, 1998.

[86] M. Mansoorizadeh and N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimed. Tools Appl.*, vol. 49, no. 2, pp. 277–297, 2010.

[87] X. Mao, L. Chen, and L. Fu, "Multi-level Speech Emotion Recognition Based on HMM and ANN," in *2009 WRI World Congress on Computer Science and Information Engineering*, 2009, vol. 7, pp. 225–229.

[88] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE ' 05 Audio-Visual Emotion Database," no. 1, 2006.

[89] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. Dagan Feng, "Video

summarization via minimum sparse reconstruction," *Pattern Recognit.*, vol. 48, no. 2, pp. 522–533, Feb. 2015.

[90] Minddisorders.com, "Affect." [Online]. Available: http://www.minddisorders.com/A-Br/Affect.html#ixzz3Zzz8UVPs.

[91] Mingliang Xue, Wanquan Liu, and Ling Li, "Person-independent facial expression recognition via hierarchical classification," in *2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, 2013, pp. 449–454.

[92] I. Mohino-Herranz, R. Gil-Pita, S. Alonso-Diaz, and M. Rosa-Zurera, "MFCC based Enlargement of the Training Set for Emotion Recognition in Speech," vol. 5, no. 1, pp. 29–40, Mar. 2014.

[93] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proceedings of the 14th ACM international conference on Multimodal interaction - ICMI '12*, 2012, p. 501.

[94] S. Nikitidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Subclass discriminant Nonnegative Matrix Factorization for facial image analysis," *Pattern Recognit.*, vol. 45, no. 12, pp. 4080–4091, Dec. 2012.

[95] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 116–125, 2012.

[96] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[97] V. Ojansivu and J. Heikkil, "Blur Insensitive Texture Classification Using Local Phase Quantization," *Image Signal Process.*, pp. 236–243, 2008.

[98] O. Onder, S. Zhalehpour, and C. E. Erdem, "A Turkish audio-visual emotional database," in *2013 21st Signal Processing and Communications Applications Conference (SIU)*, 2013, pp. 1–4.

[99] M. Paleari and B. Huet, "Toward emotion indexing of multimedia excerpts," in *2008 International Workshop on Content-Based Multimedia Indexing*, 2008, pp. 425–432.

[100] M. Paleari and C. L. Lisetti, "Toward multimodal fusion of affective cues," *HCM '06 1st ACM Int. Work. Human-centered Multimed.* -, p. 99, 2006.

[101] M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image Vis. Comput.*, vol. 18, no. 11, pp. 881–905, 2000.

[102] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-Based Database for Facial Expression Analysis," in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 317–321.

[103] M. Peura and J. Iivarinen, "Efficiency of simple shape descriptors," *Proc. Third Int. Work. Vis. Form*, pp. 443–451, 1997.

[104] T. Pfister, Xiaobai Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," in *2011 International Conference on Computer Vision*,

2011, pp. 1449–1456.

[105] Quan-You Zhao, Bao-Chang Pan, Jian-Jia Pan, and Yuan-Yan Tang, “Facial expression recognition based on fusion of Gabor and LBP features,” in *2008 International Conference on Wavelet Analysis and Pattern Recognition*, 2008, vol. 1, pp. 362–367.

[106] O. Rudovic, V. Pavlovic, and M. Pantic, “Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation,” *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. pp. 2634–2641, 2012.

[107] J. a Russell, J.-A. Bachorowski, and J.-M. Fernández-Dols, “Facial and Vocal Expressions of Emotion,” *Annu. Rev. Psychol.*, vol. 54, no. 1, pp. 329–349, Feb. 2003.

[108] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. De la Torre, and A. Rossi, “Automated Facial Expression Recognition System,” in *43rd Annual 2009 International Carnahan Conference on Security Technology*, 2009, pp. 172–177.

[109] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, “Audiovisual Behavior Modeling by Combined Feature Spaces,” *Acoust. Speech Signal Process. 2007. ICASSP 2007. IEEE Int. Conf.*, vol. 2, pp. II–733–II–736, 2007.

[110] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03).*, 2003, vol. 2, pp. II–1–4.

[111] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, “Towards more reality in the recognition of emotional speech,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 4, no. 101, pp. 941–944, 2007.

[112] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, “Facial action recognition combining heterogeneous features via multikernel learning,” *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 42, no. 4, pp. 993–1005, 2012.

[113] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on Local Binary Patterns: A comprehensive study,” *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[114] S. Shojaeilangari, “Person independent facial expression analysis using Gabor features and Genetic Algorithm,” *2011 8th Int. Conf. Information, Commun. Signal Process.*, pp. 1–5, 2011.

[115] L. C. De Silva and Pei Chi Ng, “Bimodal emotion recognition,” in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 332–335.

[116] T. Sim, S. Baker, and M. Bsat, “The CMU Pose, Illumination, and Expression (PIE) database,” *Proc. - 5th IEEE Int. Conf. Autom. Face Gesture Recognition, FGR 2002*, no. 1, pp. 53–58, 2002.

[117] D. M. Sloan and A. M. Kring, “Measuring Changes in Emotion During Psychotherapy: Conceptual and Methodological Issues,” *Clin. Psychol. Sci.*

Pract., vol. 14, no. 4, pp. 307–322, Dec. 2007.

[118] I. Sneddon, M. Mcrorie, G. Mckeown, and J. Hanratty, "The Belfast induced natural emotion database," vol. 3, no. 1, pp. 32–41, 2012.

[119] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, "'Of All Things the Measure Is Man' : Automatic Classification of Emotions and Inter-Labeler Consistency," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 2005, vol. 1, pp. 317–320.

[120] X. Sun, H. Xu, C. Zhao, and J. Yang, "Facial expression recognition based on histogram sequence of local Gabor binary patterns," *Cybernetics and Intelligent Systems, 2008 IEEE Conference on*. pp. 158–163, 2008.

[121] Y. Tong, W. Liao, and Q. Ji, "Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1683–1699, Oct. 2007.

[122] S. Ulukaya and C. E. Erdem, "Estimation of the neutral face shape using Gaussian Mixture Models," *Acoust. Speech Signal Process. (ICASSP), 2012 IEEE Int. Conf.*, pp. 1385–1388, 2012.

[123] S. Ulukaya and C. E. Erdem, "Gaussian mixture model based estimation of the neutral face shape for emotion recognition," *Digit. Signal Process.*, vol. 32, pp. 11–23, 2014.

[124] M. F. Valstar and M. Pantic, "Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database," *Proc. Int'l Conf. Lang. Resour. Eval. Work. Emot.*, pp. 65–70, 2010.

[125] M. F. Valstar, T. Almaev, J. M. Girard, G. Mckeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, "FERA 2015 - Second Facial Expression Recognition and Analysis Challenge," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015, pp. 1–8.

[126] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2729–2736.

[127] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014 - 3D Dimensional Affect and Depression Recognition Challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14*, 2014, pp. 3–10.

[128] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[129] T. Vogt and E. Andre, "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition," in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 474–477.

[130] D. Vukadinovic and M. Pantic, "Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, pp. 1692–1698.

[131] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garafolo, L. Hirschman, a. Le, S. Lee, S. S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P.

Prabhu, A. I. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker, "DARPA Communicator Dialog Travel Planning Systems : The June 2000 Data Collection," *Proc. Eurospeech 2001*, no. June, pp. 1371–1374, 2001.

[132] F. Wallhoff, B. Schuller, M. Hawellek, and G. Rigoll, "Efficient recognition of authentic dynamic facial expressions on the feedtum database," *2006 IEEE Int. Conf. Multimed. Expo, ICME 2006 - Proc.*, vol. 2006, pp. 493–496, 2006.

[133] Y. Wang, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *... , IEEE Trans.*, vol. 14, no. 3, pp. 597–607, 2012.

[134] Y. Wang and L. Guan, "Recognizing Human Emotional State From Audiovisual Signals," *IEEE Trans. Multimed.*, vol. 10, no. 5, pp. 936–946, Aug. 2008.

[135] S. Yang and B. Bhanu, "Facial expression recognition using emotion avatar image," in *Face and Gesture 2011*, 2011, pp. 866–871.

[136] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Trans. Multimed.*, vol. 8, no. 3, pp. 500–507, 2006.

[137] Yunfeng Zhu, F. De la Torre, J. F. Cohn, and Yu-Jin Zhang, "Dynamic Cascades with Bidirectional Bootstrapping for Action Unit Detection in Spontaneous Facial Behavior," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 79–91, Apr. 2011.

[138] B. Zellner-Keller, "Prediction of temporal structures for various speech rates," *Vol. Speech Synth. Springer-Verlag*, 1998.

[139] S. Zhalehpour, "Audio-Visual Affect Recognition," M.Sc. Thesis, Bahçeşehir University, 2014.

[140] S. Zhalehpour, Z. Akhtar, and C. Eroglu Erdem, "Multimodal emotion recognition with automatic peak frame selection," in *2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings*, 2014, pp. 116–121.

[141] S. Zhalehpour, Z. Akhtar, and C. Eroglu Erdem, "Multimodal emotion recognition based on peak frame selection from video," *Signal, Image Video Process.*, pp. 1–8, 2015.

[142] L. Zhang and D. Tjondronegoro, "Facial Expression Recognition Using Facial Movement Features," *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 219–229, Oct. 2011.

[143] T. Zhang, M. Hasegawa-Johnson, and S. E. Levinson, "Children's emotion recognition in an intelligent tutoring scenario," *Conf. Spok.*, pp. 1–4, 2004.

[144] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 699–714, 2005.

[145] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[146] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech Emotion Recognition Using

Both Spectral and Prosodic Features," *Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on*. pp. 1–4, 2009.

[147] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark estimation in the wild," *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 2879–2886, 2012.

[148] "http://en.wikipedia.org/wiki/Facial_Action_Coding_System." [Online]. Available: http://en.wikipedia.org/wiki/Facial_Action_Coding_System.

[149] "http://www.mimik-lesen.com/mimik-lesen-buchung.html." [Online]. Available: http://www.mimik-lesen.com/mimik-lesen-buchung.html.

[150] "Matlab codes for Local Phase Quantizatio," 2015. [Online]. Available: http://www.cse.oulu.fi/CMV/Downloads/LPQMatlab. [Accessed: 01-Jun-2015].